The genetic and ecological landscape of plasmids in the human gut

Authors: Michael K. Yu[1,‡,*], Emily C. Fogarty[2,3,‡], A. Murat Eren[2,3,4,*]

Affiliations:

[1] Toyota Technological Institute at Chicago; Chicago, IL 60637, USA

[2] Department of Medicine, University of Chicago; Chicago, IL 60637, USA

[3] Graduate Program in the Biological Sciences, The University of Chicago; Chicago, IL 60637, USA

[4] Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory; Woods Hole, MA 02543, USA

‡ These authors contributed equally.

[*]Correspondence: mikeyu@ttic.edu, meren@uchicago.edu

Abstract:

Plasmids are mobile genetic elements that often carry key determinants of fitness, yet their diversity in natural systems is poorly understood. Here we trained a machine learning model on reference genomes to recognize the genetic architecture of plasmids. We applied our model to a global collection of human gut metagenomes to identify 68,350 non-redundant plasmids, which represent a 200-fold increase in the number of detectable plasmids in the human gut. This broad view of plasmid diversity revealed 1,169 'plasmid systems', an evolutionary phenomenon where a backbone sequence with core plasmid functions, such as replication, is recombined with cargo functions, such as antibiotic resistance, depending on the environment. This work unearths the astonishing diversity of plasmids and provides a framework for studying their ecology and evolution.

One-Sentence Summary:

New computational methods enable a global search and reveal evolutionary principles of plasmids in the human gut microbiome.

Main Text:

Plasmids are a type of mobile genetic element (*1*) that occur in all domains of life (*2*, *3*). They typically exist as extrachromosomal and circular DNA, replicate semi-independently of their hosts, and often encode key determinants of fitness (*4*–*6*). Although plasmids are extensively studied due to their role in the spread of antibiotic resistance genes (*7*, *8*) and virulence factors (*9*, *10*), their significance in microbial adaptation and response to environmental change has only recently been appreciated (*11*–*13*).

Despite their significant influence on microbial phenotypes, our understanding of the diversity, ecology, and genetic architecture of plasmids is largely limited to those that are associated with cultivated organisms and is far from complete (*14*). Recent advances in technology have enabled unprecedented access to microbial genomes and metagenomes from all habitats (*15*), including the human gut (*16*, *17*), where microbes have been associated with a myriad of health and disease states (*18*). Novel sequence collections that emerge from metagenomic assemblies also have the potential to offer insights into the diversity and gene pool of plasmids; however, de novo identification of plasmids is challenging (*19*).

Several computational strategies have been developed to identify plasmids in sequence collections (*20*). Existing approaches typically search for k-mer patterns learned from known plasmid sequences (*21*–*23*) or identify known functions such as replication or conjugation genes (*24*, *25*). These methods can identify plasmids similar to those in public databases, yet they are limited in their ability to recognize novel plasmids. Other approaches focus on circularity of sequences during (meta)genomic assembly (*26*–*28*); however, this strategy overlooks plasmids that are linear, integrated, or found as assembly fragments, and may confuse other types of circular mobile elements for plasmids.

Here we present a collection of 68,350 non-redundant microbial plasmids that are substantially more prevalent across human gut microbiomes and more genetically diverse than known plasmids. We generated this collection using a machine learning approach that identifies plasmids based on *de novo* gene families, without relying on known functions or k-mer patterns. We demonstrate that many of our plasmids are novel and circular, and to further validate our model, we experimentally transfer a predicted plasmid from one microbial population to another. Finally, we implement a technical framework to define ‘plasmid systems’, which empowered us to

associate the evolution of individual plasmids (e.g. the acquisition of cargo genes like antibiotic resistance) with their ecological distribution across microbiomes.

A plasmid classification system based on *de novo* gene families

To enable a systematic study of plasmid sequences for machine learning, we compiled a reference set of 16,827 plasmids and 14,367 chromosomal sequences from public databases (Fig. 1A, table S1). In these sequences, we identified 51.2 million open reading frames and annotated them with functions defined in the Cluster of Orthologous Genes (COG) (*29*) and Pfam (*30*) databases. We also used MMseqs2 (*31*) to organize genes *de novo* into 2,322,750 gene families and removed those that contained only one gene. Of the remaining 1,090,132 *de novo* families, only 63.4% of them overlapped with COG or Pfam functions. However, *de novo* families enabled a more comprehensive analysis by accounting for 95% of all plasmid genes (Fig. 1B, fig. S1A). Using *de novo* gene families also substantially increased our ability to identify genes that were enriched in plasmids, independently of available gene function databases (fig. S1B-C).

We used this reference database to train a machine learning model, PlasX, that distinguishes between plasmids and chromosomes based on genetic architecture (Fig. 1C). PlasX is a logistic regression, which assigns a positive (or negative) coefficient to gene families that are likely to originate from a plasmid (or chromosomal) sequence. The coefficients of gene families within a sequence are summed to calculate a prediction score, ranging from 0 to 1, where a score of >0.5 designates a plasmid. To improve performance, PlasX uses a technique called elastic net regularization, which identifies gene families with redundant or noisy signals and then minimizes the usage of these families by setting their coefficients equal or close to zero. Consequently, only a non-redundant and informative set of gene families can impact predictions by having coefficients far from zero (fig. S1D). For training and evaluating PlasX, we used 10kb slices of the reference sequences, to normalize for the fact that chromosomes are generally much longer than plasmids and to improve downstream application of PlasX on sequence collections that may contain a large number of fragmented sequences, such as assembled metagenomes.

We compared PlasX to PlasClass (*23*), a state-of-the-art algorithm for plasmid classification, which also fits a logistic regression but uses k-mers of length 3-7 instead of gene families as features. We first evaluated performance in 4-fold cross-validation, using a ‘naive’ randomized splitting of sequences into training and test data (fig. S1E). PlasX achieved a higher area under

the precision-recall curve (AUCPR=0.55) than PlasClass (AUCPR=0.45, fig. S1F). While naive splitting is a common evaluation technique, it can separate very similar sequences into training and test data, especially given the redundancy of sequences in public databases, inflating the accuracy of classification. As a more accurate benchmark, we designed an 'informed' split by clustering sequences into subtypes and then keeping all sequences in the same subtype together in either the training or test data. This approach minimized the occurrence of redundant sequences in training and test data, and thus better tests the ability to recognize novel sequences. Moreover, some subtypes have many more sequences than others, reflecting the overrepresentation of well-studied microbes in public databases, and can bias performance metrics. As a more uniform survey of the diversity in our sequences, we recalculated AUCPR by assigning normalized weights to sequences such that every subtype has equal contribution to this score (*32*). These advanced evaluation procedures revealed a greater performance divide between PlasX (weighted AUCPR=0.70) versus PlasClass (weighted AUCPR=0.17) (Fig. 1D), demonstrating the importance of using gene families, rather than strictly defined sequence features, to identify new types of plasmids that are distinct from those already present in databases.

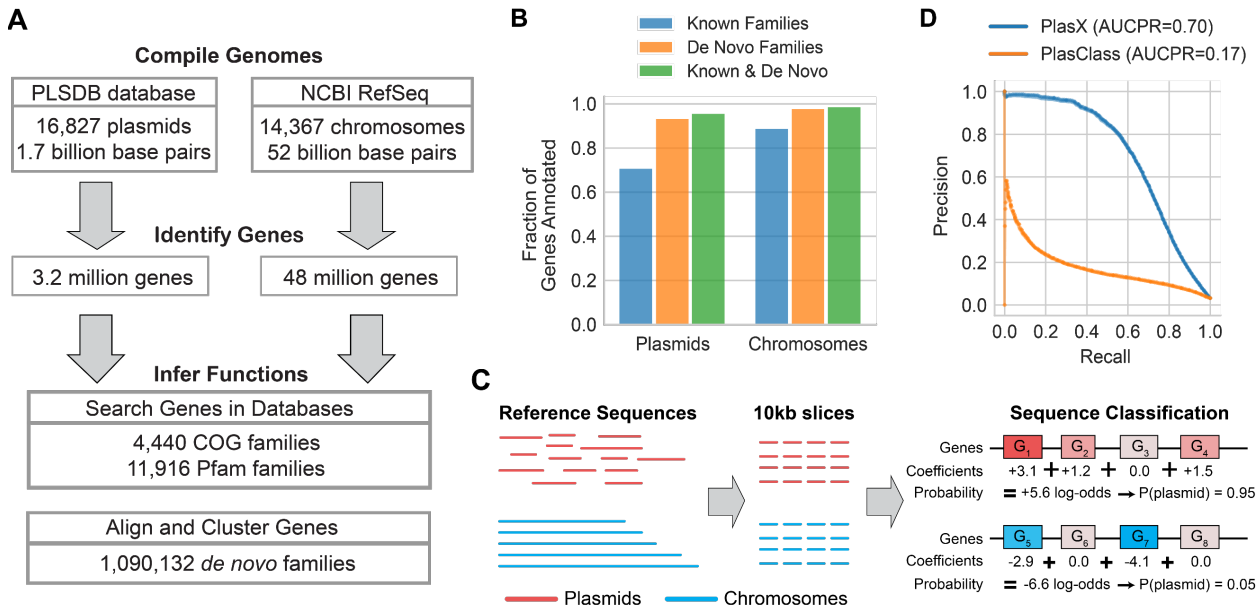

Fig. 1. A machine learning model for classifying plasmids. (A) Our pangenomics workflow to characterize gene functions in a reference set of plasmids and chromosomes. **(B)** The fraction of all plasmids or all chromosomal genes that are annotated by using known families (blue), de novo families (orange), or a combination of both (green). **(C)** Training of PlasX. Reference sequences are sliced into 10kb windows and then prediction scores are made by a logistic regression that sums the contributions of gene families within a sequence. **(D)** Precision-recall curves comparing PlasX and PlasClass, each trained using 4-fold cross-validation and an informed split. AUCPR was calculated using sequence weights for normalization.

PlasX unveils a large database of new plasmids from the human gut microbiome

Having verified PlasX's ability to identify plasmid sequences, we applied it to survey naturally occurring plasmids in the human gut microbiome, an environment which harbors a diverse range of microbes and mobile genetic elements (*33*). We assembled 36 million contigs from 1,782 human gut metagenomes, spanning culturally and geographically distinct human populations (table S2). Running PlasX on these data resulted in a total of 226,194 predicted plasmids with a model score above 0.5 (Fig. 2A, fig. S2A, table S3). While our predictions spanned a wide range of lengths, including 135 sequences that were longer than 100 kbp, they were generally shorter than known plasmids with a median length of 2.6 kbp versus 53.3 kbp, respectively (fig. S2B). This discrepancy can be partly explained by fragmentation during metagenomic assembly, as the median length of the entire set of contigs was 2.1 kbp, and only 50,310 (0.14%) contigs were longer than 100 kbp. To minimize this issue, we removed predictions that were likely fragments because they appeared to be non-circular subsequences of other predictions in our collection. This filter retained 100,719 predictions for downstream analyses.

To assess the circularity of predicted plasmids, we analyzed the orientation of paired-end reads recruited from metagenomes. This is a powerful strategy because if a contig occurred as a circular element in the environment, then paired-end reads would be recruited to opposite ends of the contig in a 'reverse-forward' orientation, instead of the typical 'forward-reverse' orientation (Fig. 2B). With this approach we found that 19,652 plasmid sequences were circular (median length: 4.4 kbp), and we designated them as high-confidence plasmids for downstream analyses. An additional 14,151 sequences were not circular themselves but were highly similar to a circular sequence. Together, these two types of sequences defined a set of 'circular-associated' sequences representing 33.6% (33,803/100,719) of predictions. Multiple factors can explain the lack of signal for circularity for the remaining plasmids, including insufficient sequencing depth to observe reverse-forward pairs, fragmented contigs, or the non-circular nature of some plasmids that occur linearly (*34*) or integrated in a chromosome (*3*).

Beyond circularity, confirming *in silico* whether a novel sequence represents a plasmid is a significant challenge. In the absence of single-copy core genes that have been vital to assess the nature of non-plasmid and non-viral genomes from metagenomes (*35*), our understanding of the canonical features of plasmids is limited to a relatively small set of well-studied genes that are

primarily derived from plasmids of model organisms in culture (*24*, *25*). For instance, MOB-suite identified canonical features for plasmid replication and conjugation in only 16.3% of the 16,827 known plasmid sequences used to train PlasX. This relatively small percentage reveals the limits of conventional approaches to identify plasmid features and thus foreshadows their limited utility to survey novel plasmids identified by advanced machine learning approaches. Indeed, MOB-suite identified canonical features in only 10.1% of our predictions. Given this narrow sensitivity, we developed orthogonal data-driven strategies to measure confidence in our predictions.

For the remaining 89.8% (90,446/100,719) of predicted plasmid sequences in which MOB-suite did not find any canonical plasmid features, we performed several types of analyses to assess how many are true plasmids or novel sequences (Fig. 2C). We found that 24.5% (24,689) of predictions were circular-associated sequences. In addition, 26.7% (26,921) were 'keyword-recognizable', as they contained a COG or Pfam function with the words 'plasmid' or 'conjugation'. And finally, 4.0% (3,996) were highly similar to a known plasmid sequence in NCBI, while 64.7% (65,117) were novel sequences with no hits to any sequence in NCBI. Combining these individual statistics, we found that 49.4% (49,739) of predictions had some orthogonal support for being a plasmid, by MOB-suite or any of the first three types of analyses, and 28.5% (28,658) had such support and were novel. Overall, these findings suggest that our collection of predicted plasmids include not only additional sequences that match known plasmids to study their diversity and gene pool in natural habitats, but also novel ones that further advance our knowledge of plasmid biology.

We further investigated the subset of predictions that were highly similar to a sequence in NCBI and categorized matches as either known plasmids (26.9%), chromosomes (21.3%), viruses (0.6%), or an unclear type of sequence (51.2%). A total of 189 predictions matched a known virus. Of these, 110 were recognized as plasmids by MOB-suite or keywords but also contained virus-related COG or Pfam functions as indicated by the keywords 'virus', 'viral', or 'phage'. These predictions carry both plasmid and viral features, a phenomenon that has previously been reported (*36*–*39*), and thus they blur the traditional categories of mobile elements. Surprisingly, 808 predictions that matched a known chromosome were also circular-associated and recognized by MOB-suite or plasmid keywords. One explanation of these data is that these plasmids can switch between an extrachromosomal or a chromosome-integrated state. Dual states are well understood in the context of synthetic vectors (*40*, *41*) but have been documented in only a small number of naturally occurring plasmids (*40*–*42*).

In general, it is likely that some of our predictions are not plasmids. To address such model errors, we observed that predictions matching known chromosomes in NCBI had lower model scores (<0.9), while those matching a known plasmid were enriched in higher model scores (>0.9) (fig. S2C). This result demonstrates that PlasX's model score is well calibrated such that a stricter threshold, such as >0.9 instead of our current choice of >0.5, could be applied to filter for more confident predictions at the cost of sensitivity (fig. S2A).

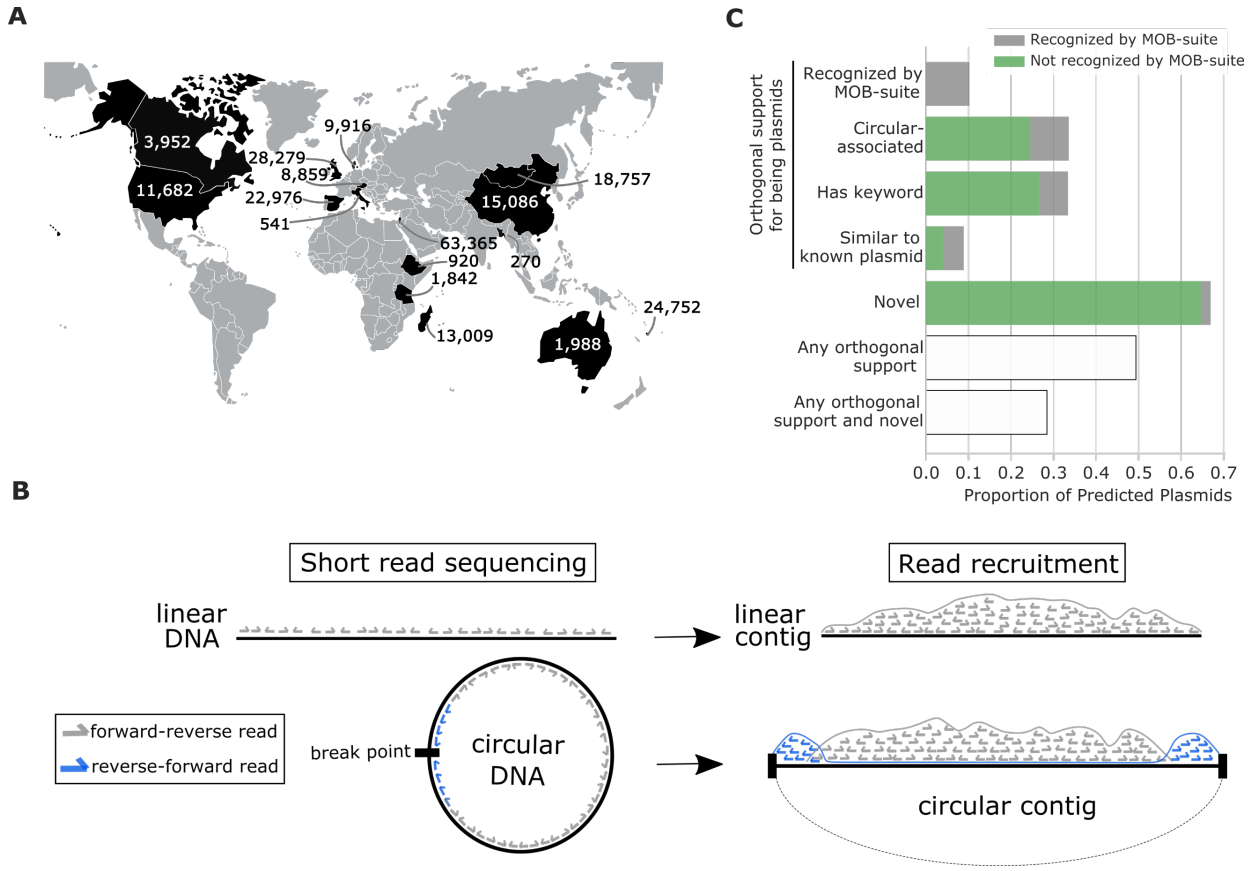

Fig. 2. Plasmid prediction from metagenomes. (A) Number of plasmids predicted from each country. **(B)** Diagram of paired-end reads mapping to a linear versus a circular contig. Linear contigs have forward-reverse reads only, while circular contigs also have reverse-forward reads concentrated on the ends due to an artifact in contig assembly. **(C)** Orthogonal support for and novelty of the 100,719 non-fragment predictions.

Plasmids predicted from metagenomes are found in isolate genomes, and capable of transfer between microbial populations

To experimentally validate our metagenome-derived predictions as true plasmids of the human gut, we developed a pipeline for identifying predictions that are (1) present in human gut microbial isolates, (2) are circular in those isolates, and (3) can be naturally transferred to other microbes. First, we detected 127 of our predicted plasmids in 14 *Bacteroides* isolate genomes that we sequenced in a previous study (*43*) (Fig. 3A). Short-read sequencing of two of these isolates suggested that the predicted plasmids pFIJ0137_1 and pENG0187_1 were circular based on paired-end orientation (Fig. 2C). We further established their circularity using long-read sequencing following a previously described approach to confirm plasmid circularity (*44*) (Fig. 3A, fig. S3). For each plasmid, we identified and manually confirmed 500 long reads that aligned completely to the plasmid but not to the host chromosome. Among these long reads, some aligned across the artificial contig breakpoint, indicating these plasmids are extrachromosomal and circular (*32*).

Finally, we tested the ability of pFIJ0137_1 to transfer between its host, *B. fragilis 214* (one of 14 isolates from (*43*)) to a well-known laboratory strain, *B. fragilis 638R.* We designed an experimental setup that takes advantage of the naturally encoded erythromycin resistance (erm^R) on pFIJ0137_1 and the rifampicin resistance (rif^R) of *B. fragilis 638R*. Specifically, we first mated isolates in the absence of antibiotics, then selected for transconjugants on media containing both antibiotics (Fig. 3B). Through short-read sequencing of the donor, recipient, and resulting transconjugants, and by employing a read recruitment analysis, we confirmed that pFIJ0137_1 transferred from *B. fragilis 214* to *B. fragilis 638R* (Fig. 3C). The read recruitment analysis also confirmed the circularity of pFIJ0137_1 in *B. fragilis 214* and both *B. fragilis 638R* transconjugants. Besides a 68bp deletion, pFIJ0137_1 in *B. fragilis* 214 (isolated in Chicago, USA) was identical to the pFIJ0137_1 version assembled from a Fijian metagenome, suggesting a relatively recent transfer of this plasmid between unrelated human populations. Overall, these experimental results show that PlasX can predict plasmids that have canonical features of being extrachromosomal, circular, and transmissible between cells.

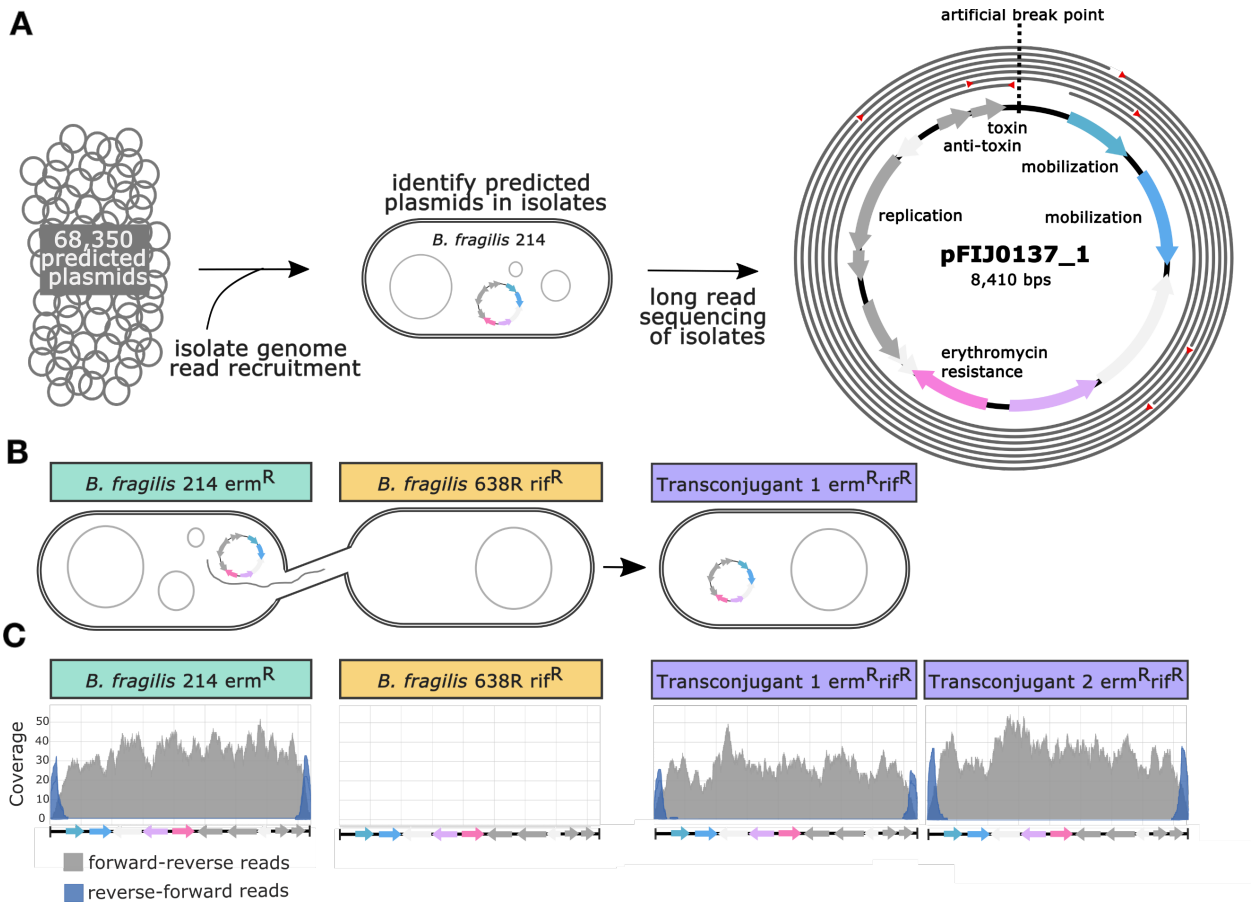

Fig. 3. Experimental validation of plasmid predictions. A) Diagram of read recruitment from 14 *Bacteroides* isolate genomes to all predicted plasmids, followed by identification and long read sequencing of an isolate, *B. fragilis* 214, containing a circular predicted plasmid (pFIJ1037_1). Grey circles represent 7 (of 500) long reads that aligned to pFIJ0137_1. Red triangles designate the beginning of a long read. **B)** Transfer of pFIJ1037_1 from *B. fragilis* 214 to *B. fragilis* 638R via conjugation and selection on erythromycin- and rifampicin-containing media. **C)** Coverage plots showing read recruitment of *B. fragilis* whole-genome sequencing reads to the pFIJ1037_1 reference sequence, confirming transfer of pFIJ1037_1. Grey are forward-reverse reads, while blue are reverse-forward reads that indicate the circularity of pFIJ1037_1.

Novel plasmids are highly prevalent, reflect human biogeography, and unexplained by microbial taxonomy

Next, we sought to characterize the ecology of plasmids across human populations through metagenomic read recruitment. For this task, we first dereplicated the entire collection of known and predicted plasmid sequences, where we assumed that any pair of plasmid sequences was redundant if at least 90% of one sequence aligned to the other with over 90% sequence identity. This analysis found 68,350 and 11,121 non-redundant sequences in the set of 226,194 predicted and 16,827 known plasmids, respectively. Then, we used the non-redundant sets of plasmids to recruit reads from the 1,782 globally distributed human gut metagenomes. We labeled a plasmid as present in a metagenome if its 'detection', the fraction of the sequence covered by at least one read, was ≥0.95 (*32*).

Our read recruitment analysis revealed that predicted plasmids were much more prevalent across human populations than known plasmids. For instance, only 1.9% (211) of known plasmids were present in at least two individuals in our dataset, suggesting the limited ecological relevance of known plasmids to naturally occurring gut microbial communities. Indeed, many of the known plasmids were isolated from a relatively small number of pathogens, such as *Escherichia coli*, *Salmonella enterica*, *Pseudomonas aeruginosa*, *Klebsiella pneumoniae*, and *Vibrio cholera*, which are unlikely to be abundant in healthy humans. In contrast, 63.1% (43,114) of the predicted plasmids were present in at least two individuals (fig. S2D). Moreover, of the most highly prevalent plasmids found in ≥100 individuals, 99.7% (5,400/5,414) were predicted plasmids while only 0.3% (14/5,414) were known. Predicted plasmids also captured the biogeography and lifestyles of human populations more effectively than known plasmids. We performed agglomerative clustering of metagenomes based on known plasmids (Fig. 4A) and found that the resulting dendrogram arranged only 50.2% of metagenomes next to another metagenome from the same country. In contrast, using predicted plasmids (Fig. 4B) resulted in 74.0% of metagenomes arranged that way. Dimensionality reduction using predicted plasmids also showed country-specific clustering (fig. S4), which was partly explained by industrialized versus non-industrialized human populations regardless of geographical distance (Fig. 4C).

These results have parallels with previous studies that found associations between the geography and lifestyles of human populations and their gut microbiota, as characterized by microbial taxonomy (*17*, *45–47*). This begs the question: Can we explain plasmid ecology simply by

taxonomy? On one hand, plasmids can only propagate within a host due to their dependence on the host's replication machinery. On the other hand, taxonomy ignores the dynamic nature of plasmids, which includes their ability to transfer across a range of taxa or their tendency to be lost or acquired by different populations of the same taxon. Here, we investigated the extent to which taxonomy explains the ecology of plasmids. For every plasmid, we identified the taxon that had the most similar ecological distribution, as scored by the correlation in abundance levels across metagenomes (*32*). We also applied thresholds to the abundance levels to determine presence or absence in a metagenome and then identified the taxon that was present in the most similar set of metagenomes as the plasmid, as scored by the Jaccard index. In most cases, the ecological similarities between predicted plasmids and their best matching taxons were low (median correlation = 0.04, median Jaccard = 0.21) (fig. S5A-B). We also observed low similarity scores even for known plasmids that are isolated from a defined microbial host (fig. S5C-D). For example, the plasmid pDOJH10S and its cognate host, *Bifidobacterium longum*, were present together in 10 metagenomes; however, 27 and 69 metagenomes contained only the plasmid or only the host, respectively (fig. S5E).

The general lack of correlation between the ecology of plasmids and microbial taxonomy in our data has implications on explaining determinants of microbial diversity in a given environment accurately. The exclusion of insights into the ecology of plasmids in microbial ecosystems may be in part responsible for the inconclusive taxonomic signal found in high-throughput analyses of human gut microbiomes that does not predict health and disease states (*48*). Despite recent advances to identify potential hosts of a given plasmid *de novo* through chromosome conformation capture (*49*), associating plasmids with their putative hosts by metagenomic sequencing remains a significant challenge. Nevertheless, our findings suggest that plasmids are a distinct feature of microbiome composition with distinct ecology that is not explained by microbial taxonomy alone.

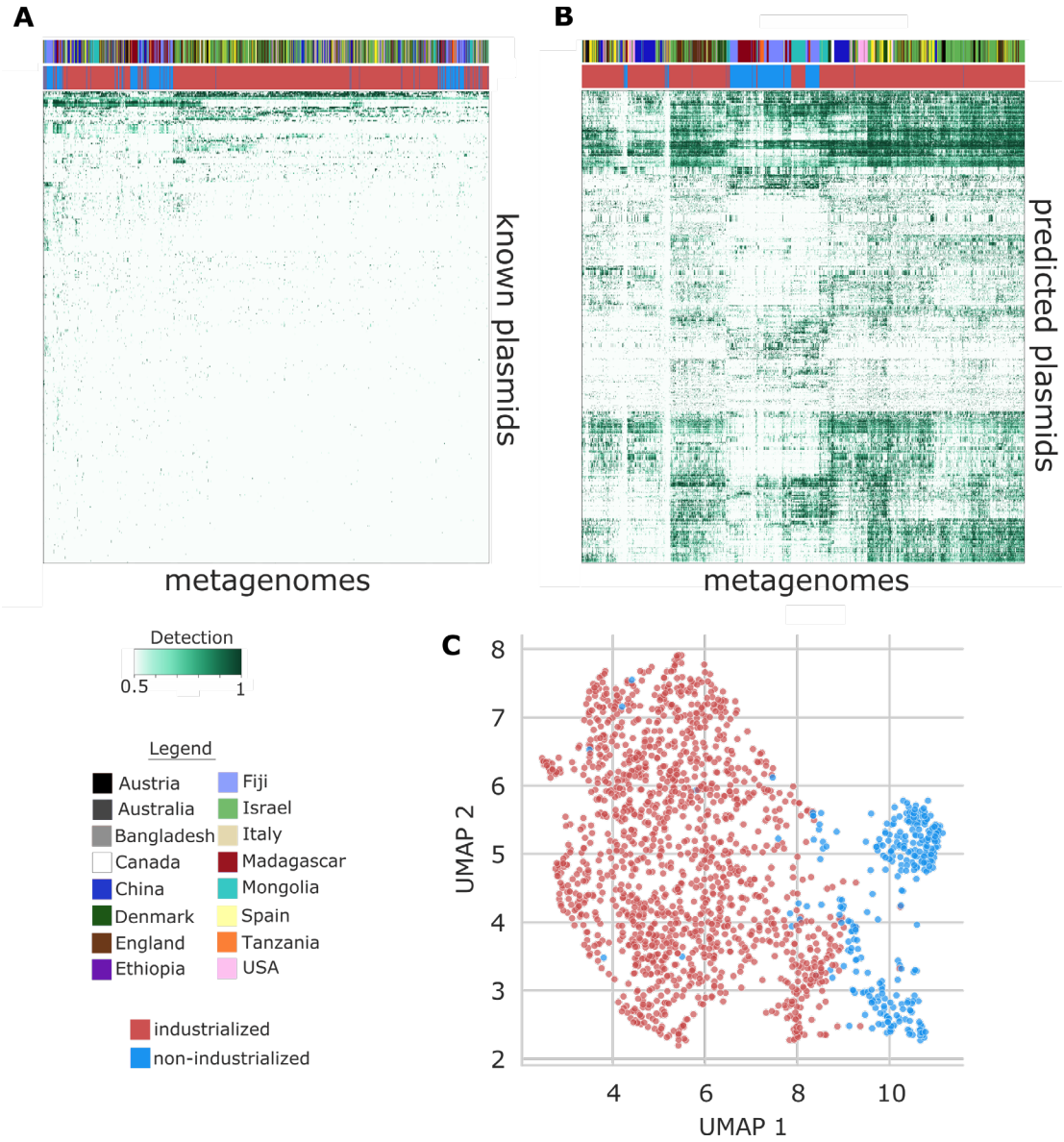

Fig. 4. Global plasmid ecology. (A) Read recruitment of human gut metagenomes to 11,121 non-redundant known plasmids. The plot shows the 338 plasmids that are present in at least one metagenome (≥0.95 detection). **(B)** Read recruitment to 68,350 non-redundant predicted plasmids. The plot shows the 1,000 most prevalent plasmids that are present in at least one metagenome and have PlasX score ≥0.75. In A and B, column colors indicate country of origin and lifestyle (industrialized or non-industrialized). **(C)** Clustering of metagenomes based on the predicted plasmids that are present, using the UMAP dimensionality reduction method (*50*). Metagenomes from industrialized or non-industrialized populations are colored red or blue, respectively.

Conserved backbones are recombined with variable cargo in plasmid systems

A hallmark of plasmids is their mosaic structure due to frequent genetic rearrangements with chromosomes and other mobile elements. While mosaicism has been observed in known plasmids (*52*) and exploited to design synthetic ones (*53*), the extent of mosaicism that occurs across closely related plasmids in naturally occurring ecosystems has not been studied. To investigate the mosaic structure of our predicted plasmids, we designed an algorithm called MobMess (Mobile Element Systems) (*32*). First, MobMess calculated pairwise alignments across all plasmids to build a sequence similarity network, in which a directed edge represents the containment of one plasmid within another (defined by ≥90% sequence identity and ≥90% coverage of the smaller plasmid) (Fig. 5A, Fig S6). Unlike many previous approaches for comparing a large collection of microbial sequences to study mosaicism (*52*), infer horizontal transfer (*54*–*56*), cluster plasmids (*57*–*59*), or cluster genomes (*60*–*62*), this network tracks the relations between smaller versus larger sequences. Next, MobMess analyzed patterns of connectivity in the network, revealing 1,169 'backbone plasmids' that each satisfy two criteria: (1) exist as a circular element, indicating the ability to propagate as an independent plasmid and (2) be encoded within a larger plasmid, indicating the ability to acquire additional genes. Each backbone plasmid represents a set of reusable genetic components (backbone genes) that acquire varying cargo genes to form 'compound plasmids'. Together, these sequences comprise a form of nested mosaicism, which we call 'plasmid systems' (Fig. 5A).

This definition of plasmid systems enables a formal categorization of plasmids into evolutionarily cohesive groups and facilitates analyses of backbone versus cargo content and their ecology, much in the same way that pangenomes enable studies of core versus accessory gene content in microbial genomes. However, plasmid systems are a specific case of pangenomics, as it is unlikely to find a naturally occurring microbial genome composed only of core genes. In contrast, backbone plasmids represent a minimal entity that can propagate using only backbone genes. Previous studies have manually identified backbone genes in well-studied plasmids (*63*, *64*), but MobMess provides an algorithmic framework and standardized vocabulary to study this concept. In addition, MobMess makes accessible a set of tunable parameters, to methodically explore the concept of plasmid systems at scale across different studies and datasets.

PlasX often assigned higher model coefficients to backbone genes in the non-redundant set of predicted plasmids, suggesting these genes define the 'essence' of a plasmid by encoding essential functions that promote the ability of a plasmid to exist as a distinct element from the chromosome, such as the genes for plasmid replication, *repL* (PF05732), and partitioning, *parA* (COG1192) (Fig. 5B). In contrast, PlasX assigned lower coefficients to cargo genes, suggesting they encode functions that are not universally essential but important for specific niches, such as nitrogen reductase, *nifH* (PF00142), and membrane transport, *ompA* (PF00691). Indeed, 24.1% (2,169/8,995) of backbone genes versus 13.1% (4,514/34,526) of cargo genes encoded COG and Pfam functions with descriptions related to plasmid replication, transfer, and maintenance (*32*).

We identified a total of 1,169 plasmid systems and named them PS1 (plasmid system #1) to PS1169. While plasmid systems captured a small fraction of the genetic diversity among non-redundant plasmids (6.5%, or 4,424/68,350), they captured a large fraction of all circular plasmid contigs (72.7%, or 14,285/19,652) (table S4). The remaining 26.9% (5,367/19,652) of circular plasmids might be part of plasmid systems in nature, but MobMess did not categorize them as so because a corresponding backbone or compound plasmid was not present in our data.

Plasmid systems were highly heterogeneous in their genetic complexity. 602 of the plasmid systems in our data contained at least 2 non-redundant compound plasmids, with the largest plasmid system containing 168 non-redundant compound plasmids (Fig. 5C). For example, pFIJ1037_1, the plasmid we isolated and transferred between *B. fragilis* organisms, was part of PS486, a system containing 24 non-redundant plasmids and found across a total of 127 metagenomes. PS486's backbone consists of a replication protein and a toxin-antitoxin system, and the cargo genes include beta-lactamases, erythromycin resistance, tetracycline resistance and riboflavin biosynthesis (Fig. 5D, table S5).

The most frequent type of function encoded on cargo genes was antibiotic resistance, including efflux pumps, which can provide general resistance to multiple antibiotics, and genes targeting specific classes of antibiotics, such as glycopeptides and beta-lactams (Fig. 6A). This large-scale observation is consistent with numerous examples of known plasmids encoding resistance and further illustrates how the widespread presence of resistance plasmids pose a public health threat (*65*, *66*).

Other highly prevalent cargo functions included a wide diversity of cellular and metabolic pathways defined in the COG (Fig. 6B) and KEGG databases (fig. S7). The most enriched among these was tRNA modification, encoded in 35 compound plasmids within 36 systems. For example, the globally prevalent system PS1110 (present in 739 metagenomes) contained 291 compound plasmids (27 non-redundant), three of which encoded an enzyme that performs tRNA Gm18 2'-O-methylation (COG0566) and were collectively present in 498 metagenomes (fig. S8, table S5). This enzyme is thought to reduce the immuno-stimulatory nature of bacterial tRNA, which is detected by Toll-like receptors (TLR7) of the mammalian innate immune system (*67*, *68*). While plasmids in some pathogens are known to facilitate bacterial evasion of mammalian immune system by regulating surface proteins (*69*), the overwhelming prevalence of tRNA modification enzymes in our data suggests the likely presence of a previously unappreciated role for plasmids to increase the fitness of their bacterial hosts against the survey of the human immune system. Distinguishing between the fundamental structure of a plasmid (backbone genes) versus the genetic currency that is exchanged (cargo genes) allowed us to organize the extensive plasmid diversity within plasmid systems and to recognize the recurrent evolution of the same cargo functions across different systems.

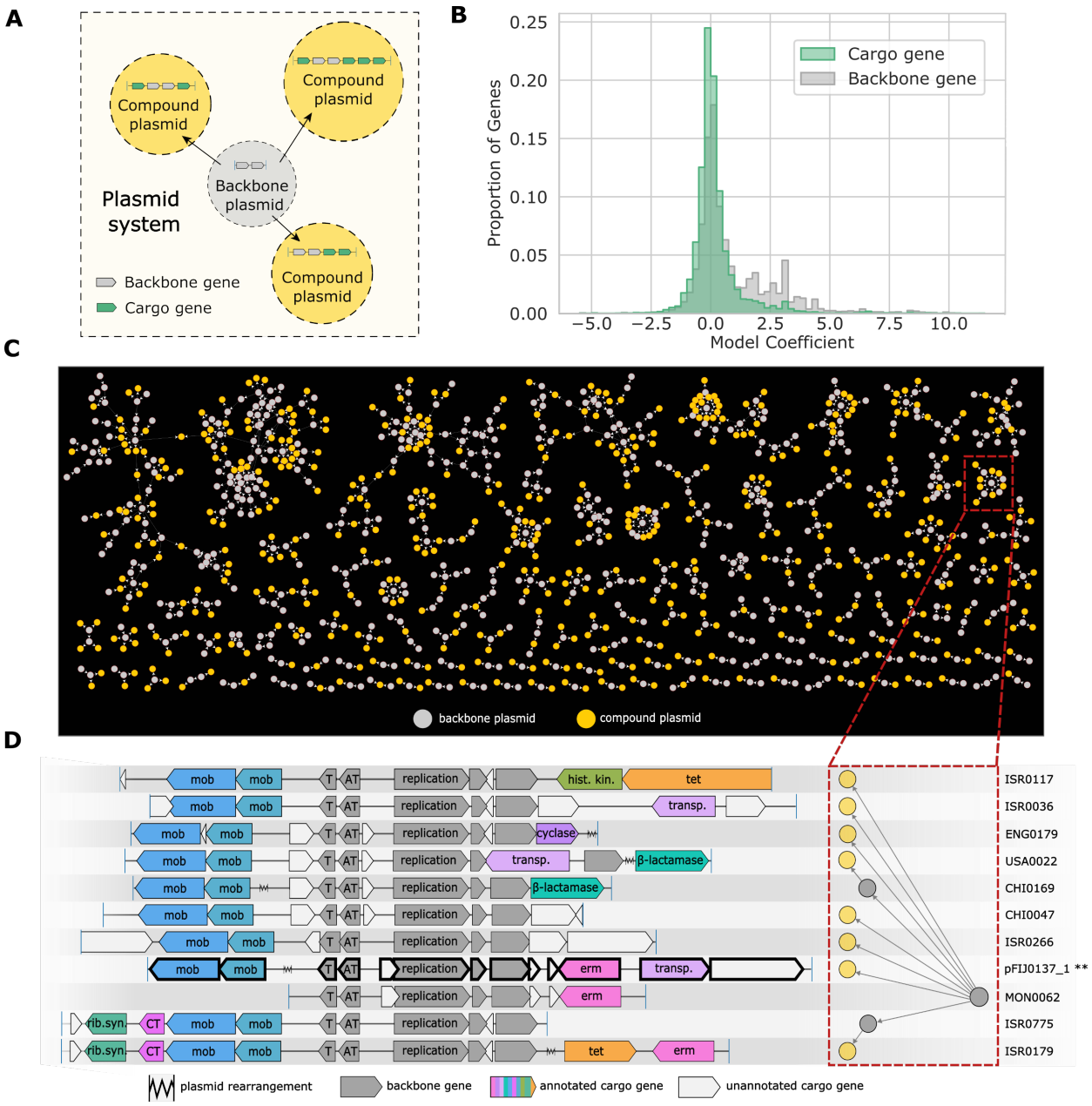

Fig. 5. **Identification of plasmid systems. (A)** Network diagram of a plasmid system. **(B)** Distribution of model coefficients for backbone vs. cargo genes in the non-redundant set of 68,350 predicted plasmids. We excluded genes that lacked gene family annotations and thus have a coefficient of zero by default. We also excluded genes that were labeled as backbone with respect to some systems but cargo in others. **(C)** Network of all plasmid systems that contain ≥3 non-redundant and high-confidence plasmids. Only these types of plasmids are shown. **(D)** Genetic architecture of plasmids in PS486, encased by a red box in C. Two plasmids in C are excluded. The system's backbone (assembled from metagenome MON0062) encodes 5 backbone genes (colored gray). Rib.syn.=riboflavin biosynthesis, CT=conjugative transfer, mob=mobilization, T=toxin, TA=anti-toxin, tet=tetracycline resistance, erm=erythromycin resistance, transp.=transposon, hist. kin.=histidine kinase.

Plasmid systems adapt their cargo genes to specific environments

We next investigated the ecological variation of plasmid systems in gut microbiomes. We found that plasmid systems were more geographically diverse than the individual plasmids in those systems (Fig. 6C, table S4). While 2,005 of these plasmids were unique to one country, 1,794 (89.5%) were part of broader systems present in at least two countries. In fact, 84.0% (982/1,169) of systems were present in at least two countries, and 9 systems were even in 15 of the 16 countries compiled in our data. We also found that plasmid systems are typically mutually exclusive, i.e. plasmids from the same system generally do not occur in the same individual human hosts. Consequently, metagenomes often have at most one plasmid from every system (fig. S9A) (*32*).

One explanation of this mutual exclusivity pattern is that a person may have never been exposed to multiple plasmids from the same system—a plausible scenario for rare systems (e.g. present in <10 metagenomes). Indeed, we observed a trend where more prevalent systems exhibit less mutual exclusivity (R^2=0.39) (fig. S9B). However, we identified many prevalent systems (e.g. present in >100 metagenomes) that are more mutually exclusive than expected. For instance, PS961 is almost perfectly mutually exclusive across 131 metagenomes (fig. S9C). For such prevalent systems, a more likely explanation lies in either the backbone or cargo content of a system. The backbones of plasmid systems typically encode genes for replication machinery, and competition for these resources within the cell can lead to incompatibility between plasmids of the same system (*70*). The large number of plasmid sequences in our dataset could support experimental investigations of the impact of incompatibility on plasmid replication (*71*, *72*), especially its contribution relative to external selective pressure on cargo genes. However, here we focused our attention on investigating ecological associations between environmental selection and cargo gene content.

The dynamic pool of cargo genes likely serves as a means for a plasmid, or its microbial host, to increase fitness in different environmental conditions. To investigate this hypothesis, we compared the evolution of plasmids in industrialized versus non-industrialized countries, an environmental difference that was reflected in plasmid content (Fig. 4C). While many plasmids systems were exclusive to one or the other type of country, 396 of them were present in both types (Fig. 6D). These global systems provide a unique *in silico* framework to explain environmental differences by variation in plasmid cargo genes.

To demonstrate this framework, we examined antibiotic usage, which can cause potent shifts in microbiome composition by enriching resistant organisms while diminishing others (*73*). In our data, the evolution of antibiotic resistance in a plasmid system coincided with the ecological variation of compound plasmids in the system. Specifically, we identified 24 high-confidence, compound plasmids that encoded antibiotic resistance in cargo genes and were exclusively present in either non-industrialized or industrialized countries (Fig. 6E). Among non-industrialized metagenomes, one of the most common types of antibiotic resistance is chloramphenicol (Fig. 6A). For instance, PS974 is highly diverse with 97 non-redundant plasmids; however, this system evolved chloramphenicol resistance (conferred via an acetyltransferase) only in plasmids assembled from Fiji (Fig. 6F, table S5). When we searched for these resistance plasmids across the global set of 1,430 metagenomes that contain PS974, we found them in 19/22 Fijian metagenomes but only in 1/1,408 non-Fijian metagenomes (p=1.1 x 10^{-13}, Fisher's exact test). Chloramphenicol is routinely prescribed in Fiji to treat eye infections, central nervous system infections, periodontitis, shigellosis, typhoid and paratyphoid fevers, and diabetic foot infections, but it is rarely used in North America and Europe (*74–77*). Thus, chloramphenicol resistance in this system likely reflects the increased exposure of Fijians to this antibiotic.

Besides chloramphenicol resistance, PS974 also contained non-Fijian compound plasmids that carry tetracycline resistance (171/1,408 metagenomes) or erythromycin resistance (429/1,408 metagenomes) (Fig. 6F). In an attempt to find an alternative explanation for the distribution of resistance plasmids in this system, we revisited our earlier question "Can plasmid ecology be simply explained by taxonomy?". By searching for these plasmids among known sequences in NCBI, we determined that possible microbial hosts include Firmicutes, such as *Blautia hydrogenotrophica*. However, none of these hosts nor any other microbial taxon had a similar ecological distribution as any of the resistance plasmids (highest Jaccard index=0.37 across all plasmid-taxon comparisons). These results suggest that compound plasmids in systems have acquired antibiotic resistance to respond to lifestyle-specific usage of antibiotics.

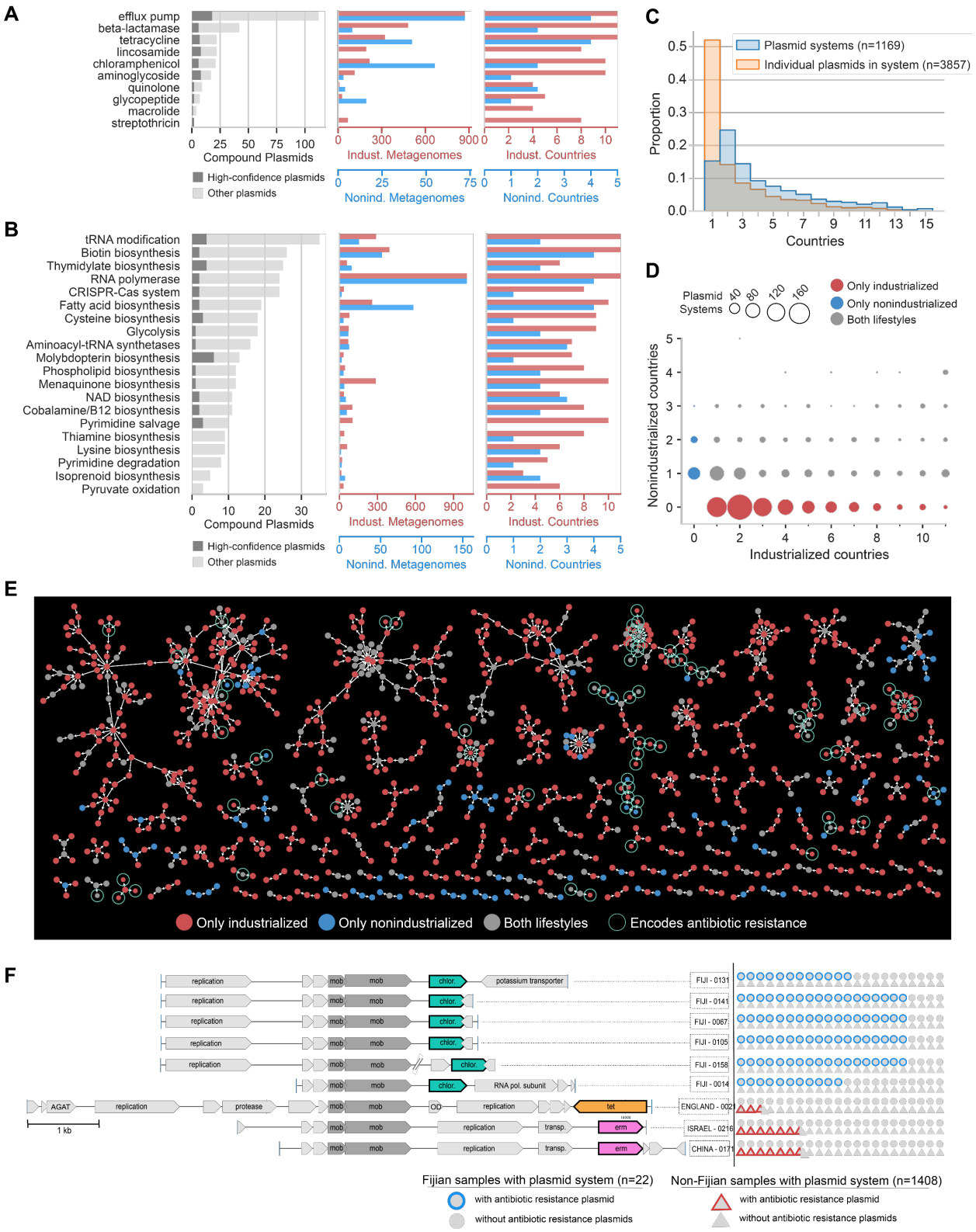

Fig. 6. Functional and ecological variation of plasmid systems. (A-B) The number of compound plasmids that encode antibiotic resistance (A) and COG pathways (B) in cargo genes. Also shown are the

metagenomes and countries containing those plasmids. In B, we show the 20 COG pathways that have the highest number of compound plasmids. To avoid redundancy with A, we exclude pathways that occur in cargo genes encoding antibiotic resistance. **(C)** Prevalence of plasmid systems versus the individual plasmids in those systems. **(D)** Distribution of plasmid systems based on the number of industrialized and non-industrialized countries they are found in. **(E)** Recoloring of the network of plasmid systems shown in Fig. 5C. Colors indicate whether a plasmid occurred in only industrialized, only non-industrialized, or both types of countries. A green ring indicates a plasmid encoding antibiotic resistance. **(F)** Compound plasmids from PS974 that encode for resistance to chloramphenicol (chlor), tetracycline (tet), or erythromycin (erm). 6/9 plasmids are circular. Dark grey genes are the backbone; light grey are cargo not related to antibiotic resistance. AGAT=aminoglycoside adenylyltransferase. OD=Oxaloacetate decarboxylase.

Discussion

Our work greatly expands the number of known plasmids by mining a global collection of metagenomes using machine learning. This expansion of the number of putative plasmid sequences has enabled us not only to conduct a comprehensive investigation of the diversity of the naturally occurring plasmid gene pool, but also propose a computational framework to partition plasmid sequences into evolutionarily cohesive units to gain deeper insights into plasmid biology. For instance, our plasmid systems revealed that naturally occurring plasmids are not static entities, but their ecology and evolution is driven by the need to respond to the environment akin to their microbial hosts. By revealing likely determinants of fitness, such as the acquisition of specific antibiotic resistance genes in Fiji as a response to a commonly used antibiotic, plasmid systems serve as a hypothesis generation and testing tool to study forces that drive plasmid evolution and influence the ecology of hosts that carry them.

Our findings suggest that new strategies to recognize plasmids, and a deliberate inclusion of additional steps for high-throughput characterization of plasmids in microbiome studies, are necessary for more complete insights into the ecology of naturally occurring microbial systems. PlasX and other plasmid recognition systems (*19*, *21–24*, *78–81*), along with MobMess to characterize plasmid systems, present a roadmap for a detailed characterization of naturally occurring plasmids, providing a complementary approach to frequently used state-of-the-art workflows to study the taxonomic composition or functional potential of environmental or host-associated microbiomes through amplicon sequences or metagenomes. Our work also complements recent efforts to characterize viruses (*82–85*) and horizontally transferred elements (*54–56*). Because we trained PlasX using a broad collection of reference sequences, it can be applied to study any environment. As PlasX recognizes plasmids using *de novo* protein families rather than known functions in curated databases, which expand much slower compared to new gene sequences, the coverage and accuracy of PlasX will increase substantially with additional plasmid sequences for training.

Plasmid systems are akin to the concept of a genetic ‘delivery van’ (backbone) with the flexibility to disseminate variable ‘packages’ (cargo genes) to microbes. This malleability is a desirable property in bioengineering and has often motivated the repurposing of naturally occurring plasmids into major tools for genetically modifying organisms. In this vein, we propose plasmid systems, especially those found to be highly prevalent and diverse, as attractive candidates for engineering tools, particularly for organisms that currently lack tools to make them genetically tractable. These systems, which manifest in many distinct forms across multiple

human populations, excel at incorporating new functions and propagating across a wide range of natural environments and thus might behave similarly in laboratory settings. During the past few decades, our ability to bypass the limitations of cultivation and study microbial genomes derived from metagenomes has led to key biological and biotechnological insights (*86*–*89*). Similarly, plasmid genomes derived from metagenomes will not only offer new opportunities to understand microbial systems, but also broaden molecular engineering tools that rely on plasmids.

Acknowledgments

We thank Karen Lolans (ORCiD:0000-0003-1903-756X) for performing the long-read sequencing and for providing feedback on the manuscript. We thank Samuel Miller (0000-0002-2836-1401) and Marcus Foo (0000-0003-3436-1632) for their insights into tRNA modification genes. We also thank other members of the Meren Lab at the University of Chicago for their feedback. MKY acknowledges support from Toyota Technological Institute at Chicago.

Funding:

Center for Data and Computing, at the University of Chicago (MKY, AME)
National Institutes of Health NIDDK grant RC2 DK122394 (AME)
Simons Foundation grant #687269 (AME)
Sloan Foundation (AME)

Author contributions:

Conceptualization: MKY, AME
Methodology: MKY, ECF, AME
Investigation: MKY, ECF, AME
Visualization: MKY, ECF, AME
Funding acquisition: MKY, AME
Project administration: MKY, AME
Supervision: MKY, AME
Writing – original draft: MKY, ECF
Writing – review & editing: MKY, ECF, AME

Data curation: MKY, ECF

Formal Analysis: MKY, ECF

Resources: ECF, MKY

Software: MKY

Validation: ECF

Competing interests:

Authors declare that they have no competing interests.

Data and materials availability:

We have released two open-source packages, PlasX (https://github.com/michaelkyu/plasx) and MobMess (https://github.com/michaelkyu/mobmess). We have also made supplementary data publicly available on zenodo.org. All analyses of reference plasmids and chromosomes are at doi:10.5281/zenodo.5732024. The PlasX model and all analyses of known and predicted plasmids are at doi:10.5281/zenodo.5732447. For all metagenomes, we have compiled the contigs, taxonomic abundances, and PlasX scores at doi:10.5281/zenodo.5730607, gene calls at doi:10.5281/zenodo.5730987, and gene annotations at doi:10.5281/zenodo.5731658. We have deposited long and short sequencing reads from *B. fragilis* isolates into the NCBI Sequence Read Archive (PRJNA782184).

Supplementary Materials and Methods

Compiling and annotating a reference set of plasmids and chromosomes

We obtained a list of 16,168 plasmids from the March 5, 2019 version of PLSDB (*90*). We also downloaded the entire collection of 13,471 complete bacterial genome assemblies from NCBI RefSeq on October 26, 2019, using instructions at https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/#allcomplete. The RefSeq assemblies contained 26,376 contigs, of which we discarded 11,350 that are also PLSDB. The reference set of 16,827 plasmids consisted of 16,168 PLSDB contigs, as well as 659 contigs from the Refseq assemblies that were labeled as 'Plasmid' in the 'Assigned-Molecule-Location/Type' field of the assembly report. The reference set of chromosomes was the remaining 14,367 Refeq contigs.

To identify and annotate genes in these sequences, we used the program `anvi-run-workflow` with `--workflow contigs` implemented (*91*) in anvi'o v7.1 (*92*), which uses Snakemake (*93*) to execute previously defined steps (https://merenlab.org/anvio-workflows/). These steps include first running Prodigal (*94*) to call genes and then running DIAMOND v2.0 (*95*) and HMMER v3.3 (*96*) on amino acid sequences to determine gene functions against the Cluster of Orthologous Groups of proteins (COGs) (*97*) and Protein Family Database models (Pfams) v32.0 (*30*), respectively. To minimize noise, we used an e-value cutoff of 10^{-10} for COGs and the default model scores for Pfams.

Modeling *de novo* gene families

We inferred *de novo* gene families by running MMseqs2 v10.6d92c (*98*) on all amino acid sequences. First, we ran `mmseqs clusthash` to collapse identical sequences into a non-redundant set for faster execution of the next step; the collapsing was inverted at the end to annotate all genes. Next, we ran `mmseqs cluster` to calculate pairwise alignments and then cluster genes that are aligned above a minimum sequence identity threshold (parameter `--min-seq-id`). We ran this program multiple times with different thresholds (0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.25, 0.2, 0.15, 0.1, 0.05) to infer a wide range of possible families. Families from different thresholds can be redundant, so we merged nested families, i.e. if family X contains all genes in family Y, then we keep X and discard Y. We also discarded any family that contains only one gene. The final result was 1,090,132 families that annotate 162,783,114 genes (a gene may have multiple annotations). This analysis took advantage of MMseqs2's parallelism, taking ~6 hours using 256 CPU cores.

Training and evaluation of PlasX and PlasClass

To infer subtypes of reference sequences, we first used mash v2.2.2 (*60*) (command `mash dist`, sketch size 100000, kmer size 21) to calculate a distance score of 0 to 1 between every pair of sequences. Next, we created an undirected graph, where sequences are nodes and sequences are connected if their distance is ≤0.1. We defined a 'subtype' as one of the 7,326 connected components in the graph. Note that a subtype can contain both plasmids and chromosomes. 3,971 subtypes contained at least one plasmid; 3,391 subtypes contained at least one chromosome.

As input data to PlasX and PlasClass, we sliced reference sequences into 10kb subsequences by sliding a 10kb window at 5kb increments. The first window starts at the beginning of the sequence, and the final window stops at the end of the sequence. For instance, a 23kb sequence would be sliced at 0-10kb, 5-15kb, 10-20kb, and 13-23kb. A slice was annotated with any gene that was entirely or partly inside the slice. We also calculated a set of weights for the 10kb slices to satisfy the following conditions: (a) the total weight of a sequence across its slices is equal to the total weight of any other sequence in the same subtype, (b) the total weight of a subtype is equal to the total weight of any other subtype, (c) the total weight of all plasmid slices is equal to that of all chromosomal slices, and (d) the total weight of all slices is equal to the number of slices. These conditions result in a unique assignment of weight values, and they are designed to minimize bias for any particular type of sequence.

We evaluated performance under two different scenarios. Scenario A uses an unweighted version of AUCPR and a 'naive' cross-validation split (fig. S1F). Scenario B uses a weighted AUCPR and an 'informed' split (Fig. 1D). We calculated AUCPR with the function *sklearn.metrics.average_precision_score* in the scikit-learn Python package (*99*). For the weighted AUCPR, we also set the function's parameter *sample_weight* to be the weights of the 10kb slices. To perform cross-validation, we randomly divided the 10kb slices into four groups. A naive split keeps slices from the same sequence together in the same group, while an informed split keeps slices from the same subtype together. In each fold of cross-validation, three groups formed the training data, and the fourth group formed the testing data.

We implemented PlasX as a logistic regression using the *SGDClassifier* class from scikit-learn. We always trained PlasX with weights, regardless of the evaluation scenario. To implement elastic net regularization, we performed a grid search of hyperparameters, with the regularization parameter *alpha* ranging from 10^{-8} to 10^{-3} in multiplicative increments of $\sqrt{10}$ and the parameter *l1_ratio* being 0.0, 0.25, 0.5, 0.75, or 1.0. For each evaluation scenario, we selected the hyperparameters that produced the best performance. We used the best hyperparameters from Scenario B (*alpha*=3.16×10^{-6}, *l1_ratio*=0.0) to retrain PlasX on all 10kb slices and create the final model that we used to predict plasmids from metagenomes.

We downloaded PlasClass v0.1 (*23*) from https://github.com/Shamir-Lab/PlasClass. We did not use weights to train PlasClass, as it does not take in weights as input. We evaluated the performances of PlasX and PlasClass using the same setup of 10kb slices, sequence weights, and cross-validation splits.

Predicting plasmids from metagenomic assemblies

We downloaded 1,782 short-read and paired-end metagenomes from the National Center for Biotechnology Information (NCBI) using the program `fastq-dump`. The countries represented are Austria (*100*), Australia (https://www.ncbi.nlm.nih.gov/bioproject/PRJEB6092), Bangladesh (*101*), Canada (*102*), China (*103*, *104*), Denmark (*105*), England (*106*), Ethiopia (*17*), Fiji (*56*), Israel (*107*), Italy (*108*), Madagascar (*17*), Mongolia (*109*), Spain (*110*), Tanzania (*108*) and USA (*111*, *112*). We labeled Tanzania, Ethiopia, Bangladesh, Madagascar, and Fiji as non-industrialized and the other countries as industrialized.

All steps of quality filtering, metagenomic assembly, read recruitment and profiling were automated using snakemake (*93*) workflows in anvi'o (*113*). The `illumina-utils` (*114*) commands `iu-gen-configs` and `iu-filter-quality-minoche` with the flag `--ignore-deflines` were used to quality

filter the raw paired-end reads. Each metagenome was assembled individually using IDBA_UD (*115*) with default settings, except the flag `--min_contig 1000`.

We annotated COGs and Pfams in all assembled contigs using the same procedure as the reference plasmids and chromosomes. To annotate *de novo* families, we first used `mmseqs result2profile` (default parameters) to represent the sequence conservation in each *de novo* family as a profile. We then used `mmseqs search` (default parameters) to search for profiles across all genes. We kept hits where the alignment coverage was ≥80% of both the gene and the profile and where the alignment identity was at least ≥X-0.05 where X is the minimum identity threshold used to infer the family (parameter --min-seq-id). For example, if a family was inferred using a threshold of 0.8, then we kept hits with an identity ≥0.75. Using these gene annotations, we ran PlasX to assign a score to every contig. We kept contigs intact, rather than slicing them into 10kb windows. Contigs with score >0.5 were classified as plasmids.

Detection and circularity of plasmids across metagenomes

We recruited short reads from our collection of metagenomes using Bowtie2 v.2.0.5 (*116*). We used the snakemake workflows in anvi'o to automate execution of bowtie and post-processing to calculate detection, the proportion of a sequence that is covered by at least one read. We ran bowtie2 using the three following combinations of parameters and input files.

First, to identify circular contigs, we recruited each metagenome's reads to a fasta file that contained only the contigs assembled from that metagenome. For computational efficiency, we ran bowtie2 with its default behavior to align every read at most once. We then analyzed the orientation of paired-end reads (Fig. 2B). During assembly, circular sequences are broken by an artificial breakpoint to represent them as linear contigs. Consequently, DNA sequencing that occurred across this breakpoint will produce paired-end reads that align in a reverse-forward orientation to the ends of the contig. In contrast, if a sequence is not circular, then all paired-end reads are expected to align in a forward-reverse orientation. To formalize this intuition, we defined a paired-end as being in a reverse-forward orientation if the upstream read, e.g. at 200-300, aligned with a reverse complement strandedness, and the downstream read, e.g. at 500-600, aligned with the same strandedness as the contig. In other words, if the contig is written 5'-to-3', then the upstream read maps 3'-to-5' and the downstream read maps 5'-to-3'. Inversely, we defined a paired-end as being in a forward-reverse orientation if the upstream read maps 5'-to-3' and the downstream read maps 3'-to-5'. Next, we defined the gap (or insert) size of a paired-end to be the distance between the farthest (or closest) aligned positions between its two reads. For example, if one read of a paired-end aligns to positions 200-300 of a contig and the other read aligns to 500-600, then the gap size is 400 and the insert size is 200. Let D be the contig's length minus three times the median insert size of forward-reverse paired-ends that are aligned to the contig. Finally, we labeled a contig as circular if (1) its detection was ≥0.95 and (2) it had at least one reverse-forward paired-end with a gap greater than or equal to D. This approach of examining reverse-forward paired-ends was inspired by (*117*).

Second, to study the ecological distribution of all plasmids and plasmid systems at the same time, we recruited each metagenome's reads to a fasta file that contained either the non-redundant set of 68,350 predicted plasmids or the non-redundant set of 11,121 known plasmids. For computational efficiency, every read was aligned at most once. We designated a plasmid as being present in a metagenome if its detection was ≥0.95. To compare metagenomes based on

their plasmid content in Fig. 4 and fig. S4, we ran UMAP v0.5.1 (*50*) with parameters `n_neighbors=30, n_components=2, min_dist=0.15, metric='jaccard', random_state=1`. The heatmaps in Fig. 4 were generated using the `heatmap.2` package in R, with agglomerative clustering using median linkage on Euclidean distances.

Third, to study the specific plasmids from PS974 and PS1110 that were shown in Fig. 6F and fig. S8 (see table S5 for contig names), we ran bowtie2 on each sequence separately. This setup allowed every read to align to potentially multiple sequences, resulting in a more complete estimation of which metagenomes contained a plasmid. For the backbone sequences of these systems, we designated them as present in a metagenome if their detection was ≥0.95. For compound plasmids in PS1110 that encoded a chloramphenicol resistance gene, we designated them as present in a metagenome if they satisfied an additional criterion that ≥0.95 of the resistance gene was covered by at least one read.

Estimation of microbial taxonomy in metagenomes

We estimated taxonomic abundances in every metagenome by running kraken2 v2.1.2 with its standard database (https://github.com/DerrickWood/kraken2) (*118*) and then refined the abundances using bracken v2.5 (https://github.com/jenniferlu717/Bracken) (*119*) with database parameters `-k 35 -l 96`. We ran bracken (parameter `-r 96`) a separate time for every taxonomic rank: S1 (subspecies/strain), S (species), G (genus), F (family), O (order), C (class), P (phylum), D (domain). The output of this analysis is a count of how many reads originated from each taxon.

To compare the metagenomic presence/absence of plasmids versus taxa, we calculated M_P, the set of metagenomes where a plasmid P is detected at ≥95%, and M_T^r, the set of metagenomes in which at least r reads originated from taxon T. For each plasmid, we attempted to find the best explanation of its ecological distribution by comparing it to every taxon using the Jaccard index, and by scanning many possible read thresholds:

$$\max_T \max_r Jaccard(P, T; r)$$

where

$$Jaccard(P, T; r) = \frac{|M_P \cap M_T^r|}{|M_P \cup M_T^r|}$$

We evaluated 29 values for the threshold r, ranging from 1 read to 10 million reads in multiplicative increments of $10^{1/4}$. We ignored plasmids that were present in less than 5 metagenomes, i.e. $|M_P|<5$, because it was likely that these plasmids would have a high Jaccard similarity to some taxon by chance. For instance, we observed many plasmids and taxa that occurred in exactly one and the same metagenome, and thus they had a Jaccard index of 1.

To compare continuous-valued abundances, we defined a plasmid's abundance in a metagenome as the sum of coverage values across all sequence positions divided by sequence length, and we defined a taxon's abundance as the number of reads originating from the taxon. If a plasmid had detection of <95%, then we set its abundance to 0. If a taxon had less than 1000 reads, then we set its abundance to 0. We ignored plasmids and taxa that had non-zero abundances in less than 5 metagenomes. For every pair of plasmid and taxon, we estimated the Pearson correlation between their abundance levels across metagenomes using FastSpar v1.0.0 (https://github.com/scwatts/fastspar) (*120*), which is an improved implementation of the SparCC method (*51*). This method accounts for the compositional nature of the data—in which

abundances reflect relative instead of absolute quantities—by assuming that the amount of correlations in a data is sparse. We ran FastSpar on the non-redundant set of predicted plasmids, and ran it again on the non-redundant set of known plasmids.

Additional annotations of plasmids

We installed MOB-suite v3.0.1 (*25*) using pip, in an Anaconda Python environment that has mash v2.2. We ran the MOB_typer subroutine (command `mob_typer`) using default parameters and followed the execution instructions at https://github.com/phac-nml/mob-suite.

Plasmids were blasted against NCBI using the *blast* package (v2.9.0, installed from bioconda Anaconda repository). On October 13, 2021, we downloaded version 5 of the NCBI databases *non-redundant nucleotide* (*nt*), *ref_prok_rep_genomes*, *ref_viroids_rep_genomes*, and *ref_viruses_rep_genomes*, and then integrated them into a single database using the `blastdb_aliastool` command. We then searched every predicted plasmids against this combined database, using the `blastn` tool with the '-task megablast' parameter for efficient searching. For each plasmid, we examined all matching NCBI sequences (called 'subjects') and chose the one with the highest qcovs (query coverage per subject), which represents the fraction of the plasmid sequence that is covered by all high-scoring segment pairs (HSP). Tiebreaking was done by sorting subjects by the maximum bitscore of the HSPs. If the qcovs of the best matching sequence was ≥90%, then we considered the predicted plasmid as found in NCBI and further categorized the matching sequence by searching for the keywords 'plasmid', 'virus', 'chromosome' (in that order, disregarding capitalization) in its NCBI description. For example, if the description of the matching sequence contained 'plasmid', then we said the predicted plasmid matched a known plasmid on NCBI. Similarly, if the description contained 'chromosome' but not 'plasmid' nor 'virus', then we said that the predicted plasmid matched a known chromosome on NCBI. If the qcovs of the best sequence was <90%, then we labeled the predicted plasmid as not found in NCBI.

To visualize pFIJ1037_1 (Fig. 3A) and pENG0187_1 (fig. S3A), we manually imported COG functions into the plasmid maps produced by snapgene (Insightful Science; snapgene.com). We manually curated functions in genes without COGs using NCBI BLASTx.

MobMess algorithm to dereplicate plasmids and discover plasmid systems (see Fig. S6)

MobMess first performs an all-vs-all pairwise alignment of sequences using the MUMmer alignment package (v4.0.0rc1) (*61*). All sequences are placed into a single fasta file and then aligned with `nucmer` (parameters `--maxmatch --minmatch=16`) to calculate local alignment blocks. Alignments are specified asymmetrically such that one sequence is designated as the query *q* and the other is the reference *r*. For every *q* and *r*, the alignment blocks calculated by `nucmer` are written to a separate file, and then a subset of blocks is identified using `delta-filter` (parameters `-q -r`) to create a one-to-one alignment.

Next, MobMess constructs a directed graph *G* where vertices are sequences and edges represent the containment of one sequence within another (fig. S6A). Formally, consider a query *q* and reference *r*. Let |*q*| be the length of *q*. For the ith alignment block between *q* and *r*, let s^i, e^i, and δ^i be the start position in *q*, end position in *q*, and number of alignment mismatches and indels, respectively. The following values summarize the information across all alignment blocks between *q* and *r*.

Sum of block lengths

$$L = \sum_i e^i - s^i$$

Number of mismatches and indels

$$E = \sum_i \delta^i$$

Proportion of query positions covered

$$C = \sum_{j=1}^{|q|} \begin{cases} 1/|q| & \text{if } \exists i \text{ such that } s^i \leq j \leq e^i \\ 0 & \text{otherwise} \end{cases}$$

Local sequence identity

$$I_{local} = (L - E)/L$$

Global sequence identity

$$I_{global} = I_{local} * C/|q|$$

MobMess creates a directed edge (*q*,*r*) in *G* if $I_{local} \geq 0.9$ and $C \geq 0.9$.

MobMess clusters sequences according to strongly connected components in *G*, calculated with igraph v0.8.2 (*121*) in Python. That is, two sequences *x* and *y* are placed in the same cluster if there exists a directed path from *x* to *y* and another from *y* to *x* in *G*. Intuitively, a cluster represents a set of sequences that are nearly identical to each other across nearly their entire lengths. MobMess then reduces *G* to another graph *H*, called the condensation graph, by contracting every cluster of sequences into a single vertex. A directed edge (*u*,*v*) exists in *H* if and only if there are sequences $x \in u$ and $y \in v$ where edge *(x,y)* exists in *G*. Note that *H* does not have any cycles. As proof by contradiction, if there were a cycle of clusters, then those clusters would have been in the same strongly connected component in *G* and hence would have been merged into a single, larger cluster.

MobMess labels every cluster in *H* as one of the three following types: (1) a 'backbone cluster' if it has an outgoing edge and at least one of its member sequences is circular, (2) a 'fragment cluster' if it has an outgoing edge but none of its member sequences are circular, or (3) a 'maximal cluster' if it does not have any outgoing edges. Intuitively, a maximal cluster represents the longest version of a plasmid observed in the data. In contrast, a fragment is so named because it may be the assembly fragment of a maximal cluster.

To dereplicate sequences, MobMess discards all fragment clusters and then chooses a representative sequence from every maximal and backbone cluster. A cluster's representative is the sequence with the highest global sequence identity (I_{global}), averaged across the set of alignments where that sequence is the reference and other sequences in the same cluster are the queries.

MobMess defines a plasmid system as a specific backbone cluster together with its 'compound' clusters, which are the set of non-fragment clusters connected to the backbone in *H*. Thus, there is a one-to-one correspondence between backbone clusters and plasmid systems. Note that systems can be nested within each other, because backbone clusters can be connected to each other in *H*. Thus, a backbone cluster can be *the backbone* that forms a given plasmid system, and at the same time, it can also be a compound cluster with respect to a smaller backbone.

We ran MobMess to analyze the 226,164 predicted plasmid sequences, yielding 132,616 clusters. Of these, 68,350 were non-fragment clusters, containing 100,719 contigs. We ran MobMess separately on the 16,827 known plasmid sequences, yielding 11,121 clusters. We assumed that all known plasmids were circular, and thus there were no fragment clusters. We visualized networks with Cytoscape v3.8 (*122*) and laid nodes out using the prefuse directed force layout (*123*). While we have focused on plasmids, MobMess could be applied to dereplicate and organize other mobile genetic elements into systems.

Classification of cargo and backbone genes

We classified genes in a given plasmid system as being backbone or cargo. Backbone genes are those encoded in any backbone plasmid of the system. Cargo genes are those that are (a) encoded in any compound plasmid of the system and (b) do not share *de novo* family annotations with any backbone genes. For *de novo* family annotations, we considered the MMseqs2 families that were derived from reference sequences to train and run PlasX, and we also generated another set of families by running mmseqs (--min-seq-id 0.05) on the set of genes from all plasmid sequences in this study (16,827 known and 226,194 predicted). This additional set of families allowed us to analyze new gene families that are only present in the predicted plasmids but not in any reference sequence. After classifying backbone and cargo genes, we labeled them as being related to plasmid replication, transfer, or maintenance by searching the description of their COGs and Pfams for any the following keywords as substrings: 'toxin', 'replicat', 'integrase', 'transpos', 'recombinase', 'resolvase', 'relaxase', 'recombination', 'partitioning', 'mobilis', 'mobiliz', 'type iv', 'conjugal', 'conjugat', 'segregat', 'MobA', 'ParA', 'ParB', 'BcsQ'.

Mutual exclusivity of plasmid systems

To measure the extent of mutual exclusivity in a plasmid system *S*, we defined two sets of metagenomes. M_s^{any} is the set of metagenomes that contains *S*, i.e. where one or more plasmids in *S* has detection at ≥0.95. M_s^{solo} is the set of metagenomes where one and only one plasmid in S has detection at ≥0.95. Then, we calculated the mutual exclusivity score E_S as the ratio $|M_s^{solo}|$ / $|M_s^{any}|$. This score is similar to a test statistic used by methods, such as CoMEt (*124*), for studying mutual exclusivity of gene alterations in cancer.

We observed that if a compound plasmid was present in a metagenome (detection ≥0.95), then its backbone plasmid was often also present. This could be due to (1) both plasmids being present as separate entities in the same microbiome, or (2) a read recruitment artifact arising from the sequence similarity between the plasmids. To ensure a non-redundant signal when calculating mutual exclusivity, we assumed the latter scenario and corrected for it by assuming backbone plasmids as absent (regardless of their detection) if any compound plasmids were present. Formally, let P_m be the set of plasmids with detection ≥0.95 in metagenome *m*. We took the induced subgraph H_m, formed by subsetting the vertices of P_m in the plasmid similarity graph *H* (see MobMess section). Then, we defined $P'_m \subseteq P_m$ as the subset of plasmids that are maximal with respect to H_m (i.e. they don't have outgoing edges in H_m). We used the plasmid-metagenome detection results implied by P'_m, rather than P_m, to calculate the mutual exclusivity score E_S.

Identification of antibiotic resistance genes

We annotated antibiotic resistance genes using two databases. First, we searched against a database of resistance protein family HMMs from Resfams (v1.2, dated 2015-01-27, 'Core' database at http://www.dantaslab.org/resfams) (*125*). We used `anvi-run-hmms` from anvi'o (*92*) to automate running `hmmsearch` from HMMER 3.3.2 (*96*) and apply an e-value cutoff of 10^{-10}. Second, we ran rgi (v5.2.0, https://github.com/arpcard/rgi) to search for similarity in the CARD database of resistance genes (*126*). We removed CARD hits that were labeled as 'Loose' and kept those labeled as 'Perfect' or 'Strict'. We removed any Resfams or CARD hits that contained the keywords 'transcription', 'regulat', 'modulat' in their database description, to avoid cases (e.g. TetR protein) where the hit is a gene that regulates the expression of another resistance gene but doesn't itself perform the molecular process that confers resistance. We categorized hits into major antibiotic resistance classes by searching for the following keywords in their functional description: 'lincosamide', 'macrolide', 'erythromycin', 'chloramphenicol', 'aminoglycoside', 'streptothricin', 'glycopeptide', 'efflux pump', 'beta-lactamase', 'nitroimidazole', 'tetraycyline', 'quinolone', 'sulfonamide'. Additionally, we searched the extra keywords 'Van' and 'VanZ' to identify glycopeptide resistance; 'efflux', 'permease', and 'pump' to identify efflux pumps; and 'TetX' to identify tetracycline resistance.

High molecular weight (HMW) DNA extraction, long-read sequencing, and determination of circularity through long-reads

We employed a long-read sequencing strategy on two *Bacteroides fragilis* cultivars from two patients (p-214 and n-216 previously described in Vineis et al (*43*)). We extracted total genomic HMW DNA by one of two methods. For *B. fragilis* p-214, we used the Qiagen Genomic Tip 20/G procedure (also known as Method #4/GT) as previously described (*127*) on a 10 mL overnight BHIS broth culture. For *B. fragilis* n-216, we used a Phenol Chloroform protocol (*127*) on 25 mL overnight BHIS broth cultures. Libraries were prepared with the Rapid Barcoding Kit (SQK-RBK004) and the standard protocols from Oxford Nanopore Technologies (ONT) with a few modifications. For *B. fragilis* p-214, DNA fragmentation was performed on 6 ug DNA using 5 passes through a 22G needle in a 30 µL volume. The gDNA input was 1.5 µg (table S6), based on sample availability in a 7.5 µL volume, with 2.5 µL Fragmentation mix added. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell (ONT). For *B. fragilis* n-216, DNA fragmentation was performed on 10 ug DNA using 10 passes through a 22G needle in a 250 µL volume. The gDNA input was 0.32 to 0.44 µg, based on sample availability in an 8.5 µL volume, with 1.5 µL Fragmentation mix was added per sample. We sequenced for 72 hours using a single R9.4/FLO-MIN106 flow cell. We used Guppy (v4.0.15) for all post-run base calling, sample de-multiplexing and the conversion of raw FAST5 to FASTQ files.

To determine circularity, we used BLAST to align the long reads with a minimum quality score of 7 to our predicted plasmid sequences. We manually examined 500 reads from each isolate to confirm (1) that they contained no chromosomal DNA, and (2) that they aligned only to our predicted plasmid. Due to DNA breakage during library preparation, each read is not necessarily the length of the plasmid. However, there were a collection of reads that overlap each other and together wrap around the entire plasmid, indicating its circularity (Fig. 3A, fig S3).

Transfer of predicted plasmid between microbial populations

In duplicate, we streaked *B. fragilis* 214 (donor, erythromycin resistant due to pFIJ0137_1) and *B. fragilis* 638R (recipient, rifampicin resistant) onto plates with brain-heart infusion agar supplemented with hemin and vitamin K (BHIS) and incubated them in 5 mL BHIS media anaerobically at 37℃ for 20 hours. To mate the donor to the recipient, 250 µL of donor cells were pelleted in a centrifuge at 5,000x gravity. We discarded the supernatant and resuspended the donor in 1 mL of the recipient culture. Again, cells were pelleted at 5,000x gravity, then resuspended in 25 µL of BHIS media. Cells were spotted onto BHIS agar plates and incubated anaerobically for 24 hours. After incubation, cells were resuspended in 1 mL BHIS. 250 µL of this suspension was plated onto BHIS plates containing 8 µg/mL rifampicin and 25 µg/mL erythromycin to select for *B. fragilis* 638R recipients of pFIJ0137_1. Duplicate plates each had approximately 300 colonies. Plating the donor or recipient alone resulted in zero colonies, confirming the transformants were not spontaneous mutants to either antibiotic. Two transformant colonies were restreaked onto fresh BHIS plates containing 8 ug/mL rifampicin and 25 µg/mL erythromycin.

Short-read sequencing of isolate genome and confirmation of plasmid transfer

We grew 20-hour cultures of *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants containing pFIJ0137_1. Using the QIAseq FX DNA library kit (Qiagen), libraries of these strains were prepared with 100 ng of genomic DNA. DNA was fragmented enzymatically into smaller fragments and desired insert size was achieved by adjusting fragmentation conditions. Fragmented DNA was end repaired and 'A's were added to the 3' ends to stage inserts for ligation. During the ligation step, Illumina compatible Unique Dual Index (UDI) adapters were added to the inserts and the prepared library was PCR amplified. Amplified libraries were cleaned up, and QC was performed using a tapestation. Libraries were sequenced on Illumina MiSeq platform using v2 cassette to generate 2x250bp reads. To confirm the transfer of pFIJ0137_1, we individually recruited reads from the *B. fragilis* 214 donor, naive *B. fragilis* 638R, and *B. fragilis* 638R transconjugants to the pFIJ0137_1 reference sequence. We used anvi'o to create contigs and profile databases (as described above) and to visualize these results with the command `anvi-interactive`. We independently confirmed the presence of pFIJ0137_1 by assembling genomes using SPAdes (*128*) with default parameters.

REFERENCES

1. L. S. Frost, R. Leplae, A. O. Summers, A. Toussaint, Mobile genetic elements: the agents of open source evolution. *Nat. Rev. Microbiol.* **3**, 722–732 (2005).

2. D. Kumar, H. K. Prajapati, A. Mahilkar, C.-H. Ma, P. Mittal, M. Jayaram, S. K. Ghosh, The selfish yeast plasmid utilizes the condensin complex and condensed chromatin for faithful partitioning. *PLoS Genet.* **17**, e1009660 (2021).

3. D. Kazlauskas, A. Varsani, E. V. Koonin, M. Krupovic, Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat. Commun.* **10**, 1–12 (2019).

4. G. del Solar, R. Giraldo, M. J. Ruiz-Echevarría, M. Espinosa, R. Díaz-Orejas, Replication and Control of Circular Bacterial Plasmids. *Microbiol. Mol. Biol. Rev.* **62**, 434 (1998).

5. S. A. Khan, Rolling-circle replication of bacterial plasmids. *Microbiology and molecular biology reviews : MMBR*. **61** (1997), pp. 442–455.

6. J. Lilly, M. Camps, Mechanisms of Theta Plasmid Replication. *Microbiol Spectr*. **3**, PLAS–0029–2014 (2015).

7. A. E. Jacob, S. J. Hobbs, Conjugal transfer of plasmid-borne multiple antibiotic resistance in Streptococcus faecalis var. zymogenes. *J. Bacteriol.* **117**, 360–372 (1974).

8. C. Poyart-Salmeron, C. Carlier, P. Trieu-Cuot, A. L. Courtieu, P. Courvalin, Transferable plasmid-mediated antibiotic resistance in Listeria monocytogenes. *Lancet*. **335**, 1422–1426 (1990).

9. R. Lan, G. Stevenson, P. R. Reeves, Comparison of Two Major Forms of the Shigella Virulence Plasmid pINV: Positive Selection Is a Major Force Driving the Divergence. *Infect. Immun.* **71**, 6298 (2003).

10. D. Meletzus, A. Bermphol, J. Dreier, R. Eichenlaub, Evidence for plasmid-encoded virulence factors in the phytopathogenic bacterium Clavibacter michiganensis subsp. michiganensis NCPPB382. *J. Bacteriol.* **175**, 2131–2136 (1993).

11. E. Broaders, C. O'Brien, C. G. M. Gahan, J. R. Marchesi, Evidence for plasmid-mediated salt tolerance in the human gut microbiome and potential mechanisms. *FEMS Microbiol. Ecol.* **92** (2016), doi:10.1093/femsec/fiw019.

12. M. Millette, C. Dupont, F. Shareck, M. T. Ruiz, D. Archambault, M. Lacroix, Purification and identification of the pediocin produced by Pediococcus acidilactici MM33, a new human intestinal strain. *J. Appl. Microbiol.* **104** (2008), doi:10.1111/j.1365-2672.2007.03583.x.

13. M. Kankainen, L. Paulin, S. Tynkkynen, I. von Ossowski, J. Reunanen, P. Partanen, R. Satokari, S. Vesterlund, A. P. A. Hendrickx, S. Lebeer, S. C. J. De Keersmaecker, J. Vanderleyden, T. Hämäläinen, S. Laukkanen, N. Salovuori, J. Ritari, E. Alatalo, R. Korpela, T. Mattila-Sandholm, A. Lassig, K. Hatakka, K. T. Kinnunen, H. Karjalainen, M. Saxelin, K. Laakso, A. Surakka, A. Palva, T. Salusjärvi, P. Auvinen, W. M. de Vos,

Comparative genomic analysis of Lactobacillus rhamnosus GG reveals pili containing a human- mucus binding protein. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17193–17198 (2009).

14. K. Smalla, S. Jechalke, E. M. Top, Plasmid detection, characterization and ecology. *Microbiology spectrum*. **3** (2015), doi:10.1128/microbiolspec.PLAS-0038-2014.

15. J. Handelsman, Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.* **68**, 669–685 (2004).

16. The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature*. **486**, 207–214 (2012).

17. E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, N. Segata, Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. *Cell*. **176**, 649–662.e20 (2019).

18. O. Manor, C. L. Dai, S. A. Kornilov, B. Smith, N. D. Price, J. C. Lovejoy, S. M. Gibbons, A. T. Magis, Health and disease markers correlate with gut microbiome composition across thousands of people. *Nat. Commun.* **11**, 1–12 (2020).

19. S. Hou, S. Cheng, T. Chen, J. A. Fuhrman, F. Sun, DeepMicrobeFinder Sorts Metagenomes into Prokaryotes, Eukaryotes and Viruses, with Marine Applications, , doi:10.21203/rs.3.rs-1016976/v1.

20. S. Arredondo-Alonso, R. J. Willems, W. van Schaik, A. C. Schürch, On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data. *Microb Genom*. **3**, e000128 (2017).

21. F. Zhou, Y. Xu, cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics*. **26**, 2051–2052 (2010).

22. P. S. Krawczyk, L. Lipinski, A. Dziembowski, PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. *Nucleic Acids Res.* **46**, e35 (2018).

23. D. Pellow, I. Mizrahi, R. Shamir, PlasClass improves plasmid sequence classification. *PLoS Comput. Biol.* **16**, e1007781 (2020).

24. A. Carattoli, E. Zankari, A. García-Fernández, M. V. Larsen, O. Lund, L. Villa, F. M. Aarestrup, H. Hasman, In SilicoDetection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. *Antimicrobial Agents and Chemotherapy*. **58** (2014), pp. 3895–3903.

25. J. Robertson, J. H. E. Nash, MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microb Genom*. **4** (2018), doi:10.1099/mgen.0.000206.

26. D. Antipov, M. Raiko, A. Lapidus, P. A. Pevzner, Plasmid detection and assembly in genomic and metagenomic data sets. *Genome Res.* **29**, 961–968 (2019).

27. R. Rozov, A. Brown Kav, D. Bogumil, N. Shterzer, E. Halperin, I. Mizrahi, R. Shamir,

Recycler: an algorithm for detecting plasmids from de novo assembly graphs. *Bioinformatics*. **33**, 475–482 (2017).

28. D. Pellow, A. Zorea, M. Probst, O. Furman, A. Segal, I. Mizrahi, R. Shamir, SCAPP: an algorithm for improved plasmid assembly in metagenomes. *Microbiome*. **9**, 144 (2021).

29. M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.* **43**, D261–9 (2015).

30. S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, M. Qureshi, L. J. Richardson, G. A. Salazar, A. Smart, E. L. L. Sonnhammer, L. Hirsh, L. Paladin, D. Piovesan, S. C. E. Tosatto, R. D. Finn, The Pfam protein families database in 2019. *Nucleic Acids Res.* **47**, D427–D432 (2019).

31. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*. **35** (2017), pp. 1026–1028.

32. Materials and methods are available as supplementary materials at the Science website.

33. V. R. Carr, A. Shkoporov, C. Hill, P. Mullany, D. L. Moyes, Probing the Mobilome: Discoveries in the Dynamic Microbiome. *Trends Microbiol.* **29**, 158–170 (2021).

34. F. Meinhardt, R. Schaffrath, M. Larsen, Microbial linear plasmids. *Appl. Microbiol. Biotechnol.* **47**, 329–336 (1997).

35. L.-X. Chen, K. Anantharaman, A. Shaiber, A. M. Eren, J. F. Banfield, Accurate and complete genomes from metagenomes. *Genome Res.* **30**, 315–333 (2020).

36. Z. Chen, L. Zhong, M. Shen, P. Fang, Z. Qin, Characterization of Streptomyces plasmid-phage pFP4 and its evolutionary implications. *Plasmid*. **68**, 170–178 (2012).

37. M. A. Oliva, A. J. Martin-Galiano, Y. Sakaguchi, J. M. Andreu, Tubulin homolog TubZ in a phage-encoded partition system. *Proc. Natl. Acad. Sci. U. S. A.* **109** (2012), doi:10.1073/pnas.1121546109.

38. T. Dokland, Molecular Piracy: Redirection of Bacteriophage Capsid Assembly by Mobile Genetic Elements. *Viruses*. **11** (2019), doi:10.3390/v11111003.

39. E. Pfeifer, J. A. Moura de Sousa, M. Touchon, E. P. C. Rocha, Bacteria have numerous distinctive groups of phage–plasmids with conserved phage and variable plasmid gene repertoires. *Nucleic Acids Res.* **49**, 2655–2673 (2021).

40. T. T. Hoang, A. J. Kutchma, A. Becher, Herbert P. Schweizer, Integration-Proficient Plasmids for Pseudomonas aeruginosa: Site-Specific Integration and Use for Engineering of Reporter and Expression Strains. *Plasmid*. **43**, 59–72 (2000).

41. P. B. Lal, F. M. Wells, Y. Lyu, I. N. Ghosh, R. Landick, P. J. Kiley, A Markerless Method for Genome Engineering in Zymomonas mobilis ZM4. *Front. Microbiol.* **0** (2019), doi:10.3389/fmicb.2019.02216.

42. K. Shishido, N. Noguchi, C. Kim, T. Ando, Isolation of a tetracycline-resistance plasmid excised from a chromosomal DNA sequence in Bacillus subtilis. *Plasmid*. **10**, 224–234

(1983).

43. J. H. Vineis, D. L. Ringus, H. G. Morrison, T. O. Delmont, S. Dalal, L. H. Raffals, D. A. Antonopoulos, D. T. Rubin, A. M. Eren, E. B. Chang, M. L. Sogin, Patient-Specific Bacteroides Genome Variants in Pouchitis. *MBio*. **7** (2016), doi:10.1128/mBio.01713-16.

44. J. Reveillaud, S. R. Bordenstein, C. Cruaud, A. Shaiber, Ö. C. Esen, M. Weill, P. Makoundou, K. Lolans, A. R. Watson, I. Rakotoarivony, S. R. Bordenstein, A. M. Eren, The Wolbachia mobilome in Culex pipiens includes a putative plasmid. *Nat. Commun.* **10**, 1051 (2019).

45. T. M. Monaghan, T. J. Sloan, S. R. Stockdale, A. M. Blanchard, R. D. Emes, M. Wilcox, R. Biswas, R. Nashine, S. Manke, J. Gandhi, P. Jain, S. Bhotmange, S. Ambalkar, A. Satav, L. A. Draper, C. Hill, R. S. Kashyap, Metagenomics reveals impact of geography and acute diarrheal disease on the Central Indian human gut microbiome. *Gut Microbes*. **12**, 1752605 (2020).

46. L. Mancabelli, C. Milani, G. A. Lugli, F. Turroni, C. Ferrario, D. van Sinderen, M. Ventura, Meta-analysis of the human gut microbiome from urbanized and pre-agricultural populations. *Environ. Microbiol.* **19**, 1379–1390 (2017).

47. T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, J. I. Gordon, Human gut microbiome viewed across age and geography. *Nature*. **486**, 222–227 (2012).

48. J. Lloyd-Price, C. Arze, A. N. Ananthakrishnan, M. Schirmer, J. Avila-Pacheco, T. W. Poon, E. Andrews, N. J. Ajami, K. S. Bonham, C. J. Brislawn, D. Casero, H. Courtney, A. Gonzalez, T. G. Graeber, A. B. Hall, K. Lake, C. J. Landers, H. Mallick, D. R. Plichta, M. Prasad, G. Rahnavard, J. Sauk, D. Shungin, Y. Vázquez-Baeza, R. A. White 3rd, IBDMDB Investigators, J. Braun, L. A. Denson, J. K. Jansson, R. Knight, S. Kugathasan, D. P. B. McGovern, J. F. Petrosino, T. S. Stappenbeck, H. S. Winter, C. B. Clish, E. A. Franzosa, H. Vlamakis, R. J. Xavier, C. Huttenhower, Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature*. **569**, 655–662 (2019).

49. M. Marbouty, A. Cournac, J.-F. Flot, H. Marie-Nelly, J. Mozziconacci, R. Koszul, Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. *Elife*. **3**, e03318 (2014).

50. L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018), (available at http://arxiv.org/abs/1802.03426).

51. J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8**, e1002687 (2012).

52. M. W. Pesesky, R. Tilley, D. A. C. Beck, Mosaic plasmids are abundant and unevenly distributed across prokaryotic taxa. *Plasmid*. **102**, 10–18 (2019).

53. J. Brosius, A. Ullrich, M. A. Raker, A. Gray, T. J. Dull, R. R. Gutell, H. F. Noller, Construction and fine mapping of recombinant plasmids containing the rrnB ribosomal

RNA operon of E. coli. *Plasmid*. **6**, 112–118 (1981).

54. C. S. Smillie, M. B. Smith, J. Friedman, O. X. Cordero, L. A. David, E. J. Alm, Ecology drives a global network of gene exchange connecting the human microbiome. *Nature*. **480**, 241–244 (2011).

55. M. Groussin, M. Poyet, A. Sistiaga, S. M. Kearney, K. Moniz, M. Noel, J. Hooker, S. M. Gibbons, L. Segurel, A. Froment, R. S. Mohamed, A. Fezeu, V. A. Juimo, S. Lafosse, F. E. Tabe, C. Girard, D. Iqaluk, L. T. T. Nguyen, B. J. Shapiro, J. Lehtimäki, L. Ruokolainen, P. P. Kettunen, T. Vatanen, S. Sigwazi, A. Mabulla, M. Domínguez-Rodrigo, Y. A. Nartey, A. Agyei-Nkansah, A. Duah, Y. A. Awuku, K. A. Valles, S. O. Asibey, M. Y. Afihene, L. R. Roberts, A. Plymoth, C. A. Onyekwere, R. E. Summons, R. J. Xavier, E. J. Alm, Elevated rates of horizontal gene transfer in the industrialized human microbiome. *Cell*. **184**, 2053–2067.e18 (2021).

56. I. L. Brito, S. Yilmaz, K. Huang, L. Xu, S. D. Jupiter, A. P. Jenkins, W. Naisilisili, M. Tamminen, C. S. Smillie, J. R. Wortman, B. W. Birren, R. J. Xavier, P. C. Blainey, A. K. Singh, D. Gevers, E. J. Alm, Mobile genes in the human microbiome are structured from global to individual scales. *Nature*. **535**, 435 (2016).

57. M. Acman, L. van Dorp, J. M. Santini, F. Balloux, Large-scale network analysis captures biological features of bacterial plasmids. *Nat. Commun.* **11**, 1–11 (2020).

58. J. Robertson, K. Bessonov, J. Schonfeld, J. H. E. Nash, Universal whole-sequence-based plasmid typing and its utility to prediction of host range and epidemiological surveillance. *Microb Genom*. **6** (2020), doi:10.1099/mgen.0.000435.

59. T. F. Jesus, B. Ribeiro-Gonçalves, D. N. Silva, V. Bortolaia, M. Ramirez, J. A. Carriço, Plasmid ATLAS: plasmid visual analytics and identification in high-throughput sequencing data. *Nucleic Acids Res.* **47**, D188–D194 (2018).

60. B. D. Ondov, T. J. Treangen, P. Melsted, A. B. Mallonee, N. H. Bergman, S. Koren, A. M. Phillippy, Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).

61. C. Jain, L. M. Rodriguez-R, A. M. Phillippy, K. T. Konstantinidis, S. Aluru, High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).

62. M. R. Olm, C. T. Brown, B. Brooks, J. F. Banfield, dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

63. C. M. Thomas, L. S. Frost, Plasmid Genomes, Introduction to. *Molecular Life Sciences* (2018), pp. 935–954.

64. C. Revilla, M. P. Garcillán-Barcia, R. Fernández-López, N. R. Thomson, M. Sanders, M. Cheung, C. M. Thomas, F. de la Cruz, Different pathways to acquiring resistance genes illustrated by the recent evolution of IncW plasmids. *Antimicrob. Agents Chemother.* **52**, 1472–1480 (2008).

65. C. O. Vrancianu, L. I. Popa, C. Bleotu, M. C. Chifiriuc, Targeting Plasmids to Limit

Acquisition and Transmission of Antimicrobial Resistance. *Front. Microbiol.* **11**, 761 (2020).

66. R. C. MacLean, A. San Millan, The evolution of antibiotic resistance. *Science*. **365**, 1082–1083 (2019).

67. S. Gehrig, M.-E. Eberle, F. Botschen, K. Rimbach, F. Eberle, T. Eigenbrod, S. Kaiser, W. M. Holmes, V. A. Erdmann, M. Sprinzl, G. Bec, G. Keith, A. H. Dalpke, M. Helm, Identification of modifications in microbial, native tRNA that suppress immunostimulatory activity. *J. Exp. Med.* **209**, 225–233 (2012).

68. A. Galvanin, L.-M. Vogt, A. Grober, I. Freund, L. Ayadi, V. Bourguignon-Igel, L. Bessler, D. Jacob, T. Eigenbrod, V. Marchand, A. Dalpke, M. Helm, Y. Motorin, Bacterial tRNA 2'-O-methylation is dynamically regulated under stress conditions and modulates innate immune response. *Nucleic Acids Res.* **48**, 12833–12844 (2020).

69. M. E. Embers, X. Alvarez, T. Ooms, M. T. Philipp, The failure of immune response evasion by linear plasmid 28-1-deficient Borrelia burgdorferi is attributable to persistent expression of an outer surface protein. *Infect. Immun.* **76**, 3984–3991 (2008).

70. C. M. Thomas, Plasmid Incompatibility. *Molecular Life Sciences* (2014), pp. 1–3.

71. R. P. Novick, Plasmid incompatibility. *Microbiological Reviews*. **51** (1987), pp. 381–395.

72. N. Velappan, D. Sblattero, L. Chasteen, P. Pavlik, A. R. M. Bradbury, Plasmid incompatibility: more compatible than previously thought? *Protein Eng. Des. Sel.* **20**, 309–313 (2007).

73. A. Palleja, K. H. Mikkelsen, S. K. Forslund, A. Kashani, K. H. Allin, T. Nielsen, T. H. Hansen, S. Liang, Q. Feng, C. Zhang, P. T. Pyl, L. P. Coelho, H. Yang, J. Wang, A. Typas, M. F. Nielsen, H. B. Nielsen, P. Bork, J. Wang, T. Vilsbøll, T. Hansen, F. K. Knop, M. Arumugam, O. Pedersen, Recovery of gut microbiota of healthy adults following antibiotic exposure. *Nature Microbiology*. **3**, 1255–1265 (2018).

74. B. Berendsen, L. Stolker, J. de Jong, M. Nielen, E. Tserendorj, R. Sodnomdarjaa, A. Cannavan, C. Elliott, Evidence of natural occurrence of the banned antibiotic chloramphenicol in herbs and grass. *Anal. Bioanal. Chem.* **397**, 1955 (2010).

75. L. Both, R. Botgros, M. Cavaleri, Analysis of licensed over-the-counter (OTC) antibiotics in the European Union and Norway, 2012. *Euro Surveill.* **20**, 30002 (2015).

76. H. J. Balbi, Chloramphenicol: A Review. *Pediatrics in Review*. **25** (2004), pp. 284–288.

77. Ministry of Health and Medical Services, Government of Fiji, "Fiji Antibiotic Guidelines" (Government of Fiji, 2019).

78. O. Schwengers, P. Barth, L. Falgenhauer, T. Hain, T. Chakraborty, A. Goesmann, Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores. *Microb Genom*. **6** (2020), doi:10.1099/mgen.0.000398.

79. G. Royer, J. W. Decousser, C. Branger, M. Dubois, C. Médigue, E. Denamur, D. Vallenet, PlaScope: a targeted approach to assess the plasmidome from genome assemblies at the

species level. *Microb Genom*. **4** (2018), doi:10.1099/mgen.0.000211.

80. S. Arredondo-Alonso, M. R. C. Rogers, J. C. Braat, T. D. Verschuuren, J. Top, J. Corander, R. J. L. Willems, A. C. Schürch, Mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species. *Microb. Genom.* **4** (2018), doi:10.1099/mgen.0.000224.

81. R. Gomi, K. L. Wyres, K. E. Holt, Detection of plasmid contigs in draft genome assemblies using customized Kraken databases. *Microb Genom*. **7** (2021), doi:10.1099/mgen.0.000550.

82. J. Guo, B. Bolduc, A. A. Zayed, A. Varsani, G. Dominguez-Huerta, T. O. Delmont, A. A. Pratama, M. C. Gazitúa, D. Vik, M. B. Sullivan, S. Roux, VirSorter2: a multi-classifier, expert-guided approach to detect diverse DNA and RNA viruses. *Microbiome*. **9**, 37 (2021).

83. B. Al-Shayeb, R. Sachdeva, L.-X. Chen, F. Ward, P. Munk, A. Devoto, C. J. Castelle, M. R. Olm, K. Bouma-Gregson, Y. Amano, C. He, R. Méheust, B. Brooks, A. Thomas, A. Lavy, P. Matheus-Carnevali, C. Sun, D. S. A. Goltsman, M. A. Borton, A. Sharrar, A. L. Jaffe, T. C. Nelson, R. Kantor, R. Keren, K. R. Lane, I. F. Farag, S. Lei, K. Finstad, R. Amundson, K. Anantharaman, J. Zhou, A. J. Probst, M. E. Power, S. G. Tringe, W.-J. Li, K. Wrighton, S. Harrison, M. Morowitz, D. A. Relman, J. A. Doudna, A.-C. Lehours, L. Warren, J. H. D. Cate, J. M. Santini, J. F. Banfield, Clades of huge phages from across Earth's ecosystems. *Nature*. **578**, 425–431 (2020).

84. A. N. Shkoporov, C. Hill, Bacteriophages of the Human Gut: The "Known Unknown" of the Microbiome. *Cell Host Microbe*. **25**, 195–209 (2019).

85. D. Antipov, M. Raiko, A. Lapidus, P. A. Pevzner, Metaviral SPAdes: assembly of viruses from metagenomic data. *Bioinformatics*. **36**, 4126–4129 (2020).

86. T. O. Delmont, C. Quince, A. Shaiber, Ö. C. Esen, S. T. Lee, M. S. Rappé, S. L. McLellan, S. Lücker, A. M. Eren, Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat Microbiol*. **3**, 804–813 (2018).

87. M. A. H. J. van Kessel, D. R. Speth, M. Albertsen, P. H. Nielsen, H. J. M. Op den Camp, B. Kartal, M. S. M. Jetten, S. Lücker, Complete nitrification by a single microorganism. *Nature*. **528**, 555–559 (2015).

88. R. A. Edwards, A. A. Vega, H. M. Norman, M. Ohaeri, K. Levi, E. A. Dinsdale, O. Cinek, R. K. Aziz, K. McNair, J. J. Barr, K. Bibby, S. J. J. Brouns, A. Cazares, P. A. de Jonge, C. Desnues, S. L. Díaz Muñoz, P. C. Fineran, A. Kurilshikov, R. Lavigne, K. Mazankova, D. T. McCarthy, F. L. Nobrega, A. Reyes Muñoz, G. Tapia, N. Trefault, A. V. Tyakht, P. Vinuesa, J. Wagemans, A. Zhernakova, F. M. Aarestrup, G. Ahmadov, A. Alassaf, J. Anton, A. Asangba, E. K. Billings, V. A. Cantu, J. M. Carlton, D. Cazares, G.-S. Cho, T. Condeff, P. Cortés, M. Cranfield, D. A. Cuevas, R. De la Iglesia, P. Decewicz, M. P. Doane, N. J. Dominy, L. Dziewit, B. M. Elwasila, A. M. Eren, C. Franz, J. Fu, C. Garcia-Aljaro, E. Ghedin, K. M. Gulino, J. M. Haggerty, S. R. Head, R. S. Hendriksen, C. Hill, H. Hyöty, E. N. Ilina, M. T. Irwin, T. C. Jeffries, J. Jofre, R. E. Junge, S. T. Kelley, M. Khan Mirzaei, M. Kowalewski, D. Kumaresan, S. R. Leigh, D. Lipson, E. S. Lisitsyna, M. Llagostera, J. M. Maritz, L. C. Marr, A. McCann, S. Molshanski-Mor, S. Monteiro, B. Moreira-Grez, M. Morris, L. Mugisha, M. Muniesa, H. Neve, N.-P. Nguyen, O. D. Nigro, A.

S. Nilsson, T. O'Connell, R. Odeh, A. Oliver, M. Piuri, A. J. Prussin Ii, U. Qimron, Z.-X. Quan, P. Rainetova, A. Ramírez-Rojas, R. Raya, K. Reasor, G. A. O. Rice, A. Rossi, R. Santos, J. Shimashita, E. N. Stachler, L. C. Stene, R. Strain, R. Stumpf, P. J. Torres, A. Twaddle, M. Ugochi Ibekwe, N. Villagra, S. Wandro, B. White, A. Whiteley, K. L. Whiteson, C. Wijmenga, M. M. Zambrano, H. Zschach, B. E. Dutilh, Global phylogeography and ancient evolution of the widespread human gut virus crAssphage. *Nat Microbiol*. **4**, 1727–1736 (2019).

89. L. A. Hug, B. J. Baker, K. Anantharaman, C. T. Brown, A. J. Probst, C. J. Castelle, C. N. Butterfield, A. W. Hernsdorf, Y. Amano, K. Ise, Y. Suzuki, N. Dudek, D. A. Relman, K. M. Finstad, R. Amundson, B. C. Thomas, J. F. Banfield, A new view of the tree of life. *Nat Microbiol*. **1**, 16048 (2016).

90. V. Galata, T. Fehlmann, C. Backes, A. Keller, PLSDB: a resource of complete bacterial plasmids. *Nucleic Acids Res.* **47**, D195–D202 (2019).

91. A. Shaiber, A. D. Willis, T. O. Delmont, S. Roux, L.-X. Chen, A. C. Schmid, M. Yousef, A. R. Watson, K. Lolans, Ö. C. Esen, S. T. M. Lee, N. Downey, H. G. Morrison, F. E. Dewhirst, J. L. Mark Welch, A. M. Eren, Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol.* **21**, 292 (2020).

92. A. M. Eren, E. Kiefl, A. Shaiber, I. Veseli, S. E. Miller, M. S. Schechter, I. Fink, J. N. Pan, M. Yousef, E. C. Fogarty, F. Trigodet, A. R. Watson, Ö. C. Esen, R. M. Moore, Q. Clayssen, M. D. Lee, V. Kivenson, E. D. Graham, B. D. Merrill, A. Karkman, D. Blankenberg, J. M. Eppley, A. Sjödin, J. J. Scott, X. Vázquez-Campos, L. J. McKay, E. A. McDaniel, S. L. R. Stevens, R. E. Anderson, J. Fuessel, A. Fernandez-Guerra, L. Maignien, T. O. Delmont, A. D. Willis, Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol*. **6**, 3–6 (2021).

93. J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*. **28**, 2520–2522 (2012).

94. D. Hyatt, G.-L. Chen, P. F. Locascio, M. L. Land, F. W. Larimer, L. J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*. **11**, 119 (2010).

95. B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*. **12**, 59–60 (2015).

96. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

97. M. Y. Galperin, Y. I. Wolf, K. S. Makarova, R. Vera Alvarez, D. Landsman, E. V. Koonin, COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* **49**, D274–D281 (2021).

98. M. Steinegger, J. Söding, Clustering huge protein sequence sets in linear time. *Nat. Commun.* **9**, 1–8 (2018).

99. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, Others, Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*. **12**, 2825–2830 (2011).

100. Q. Feng, S. Liang, H. Jia, A. Stadlmayr, L. Tang, Z. Lan, D. Zhang, H. Xia, X. Xu, Z. Jie, L. Su, X. Li, X. Li, J. Li, L. Xiao, U. Huber-Schönauer, D. Niederseer, X. Xu, J. Y. Al-Aama, H. Yang, J. Wang, K. Kristiansen, M. Arumugam, H. Tilg, C. Datz, J. Wang, Gut microbiome development along the colorectal adenoma–carcinoma sequence. *Nat. Commun.* **6**, 1–13 (2015).

101. L. A. David, A. Weil, E. T. Ryan, S. B. Calderwood, J. B. Harris, F. Chowdhury, Y. Begum, F. Qadri, R. C. LaRocque, P. J. Turnbaugh, Gut Microbial Succession Follows Acute Secretory Diarrhea in Humans. *MBio*. **6** (2015), doi:10.1128/mBio.00381-15.

102. F. Raymond, A. A. Ouameur, M. Déraspe, N. Iqbal, H. Gingras, B. Dridi, P. Leprohon, P.-L. Plante, R. Giroux, È. Bérubé, J. Frenette, D. K. Boudreau, J.-L. Simard, I. Chabot, M.-C. Domingo, S. Trottier, M. Boissinot, A. Huletsky, P. H. Roy, M. Ouellette, M. G. Bergeron, J. Corbeil, The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J.* **10**, 707 (2016).

103. J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, J. Wang, A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*. **464**, 59–65 (2010).

104. C. Wen, Z. Zheng, T. Shao, L. Liu, Z. Xie, E. Le Chatelier, Z. He, W. Zhong, Y. Fan, L. Zhang, H. Li, C. Wu, C. Hu, Q. Xu, J. Zhou, S. Cai, D. Wang, Y. Huang, M. Breban, N. Qin, S. D. Ehrlich, Quantitative metagenomics reveals unique gut microbiome biomarkers in ankylosing spondylitis. *Genome Biol.* **18** (2017), doi:10.1186/s13059-017-1271-6.

105. E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J. M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jørgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J. D. Zucker, J. Raes, T. Hansen, P. Bork, J. Wang, S. D. Ehrlich, O. Pedersen, Richness of human gut microbiome correlates with metabolic markers. *Nature*. **500** (2013), doi:10.1038/nature12506.

106. H. Xie, R. Guo, H. Zhong, Q. Feng, Z. Lan, B. Qin, K. J. Ward, M. A. Jackson, Y. Xia, X. Chen, B. Chen, H. Xia, C. Xu, F. Li, X. Xu, J. Y. Al-Aama, H. Yang, J. Wang, K. Kristiansen, J. Wang, C. J. Steves, J. T. Bell, J. Li, T. D. Spector, H. Jia, Shotgun Metagenomics of 250 Adult Twins Reveals Genetic and Environmental Impacts on the Gut Microbiome. *Cell systems*. **3**, 572 (2016).

107. D. Zeevi, T. Korem, N. Zmora, D. Israeli, D. Rothschild, A. Weinberger, O. Ben-Yacov, D. Lador, T. Avnit-Sagi, M. Lotan-Pompan, J. Suez, J. A. Mahdi, E. Matot, G. Malka, N. Kosower, M. Rein, G. Zilberman-Schapira, L. Dohnalová, M. Pevsner-Fischer, R. Bikovsky, Z. Halpern, E. Elinav, E. Segal, Personalized Nutrition by Prediction of Glycemic Responses. *Cell*. **163**, 1079–1094 (2015).

108. S. Rampelli, S. L. Schnorr, C. Consolandi, S. Turroni, M. Severgnini, C. Peano, P. Brigidi,

A. N. Crittenden, A. G. Henry, M. Candela, Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. *Curr. Biol.* **25**, 1682–1693 (2015).

109. W. Liu, J. Zhang, C. Wu, S. Cai, W. Huang, J. Chen, X. I. Xiaoxia, Z. Liang, Q. Hou, B. Zhou, N. Qin, H. Zhang, Unique Features of Ethnic Mongolian Gut Microbiome revealed by metagenomic analysis. *Sci. Rep.* **6** (2016), doi:10.1038/srep34826.

110. J. Li, H. Jia, X. Cai, H. Zhong, Q. Feng, S. Sunagawa, M. Arumugam, J. R. Kultima, E. Prifti, T. Nielsen, A. S. Juncker, C. Manichanh, B. Chen, W. Zhang, F. Levenez, J. Wang, X. Xu, L. Xiao, S. Liang, D. Zhang, Z. Zhang, W. Chen, H. Zhao, J. Y. Al-Aama, S. Edris, H. Yang, J. Wang, T. Hansen, H. B. Nielsen, S. Brunak, K. Kristiansen, F. Guarner, O. Pedersen, J. Doré, S. D. Ehrlich, P. Bork, J. Wang, An integrated catalog of reference genes in the human gut microbiome. *Nat. Biotechnol.* **32**, 834–841 (2014).

111. P. J. Turnbaugh, R. E. Ley, M. Hamady, C. M. Fraser-Liggett, R. Knight, J. I. Gordon, The Human Microbiome Project. *Nature*. **449**, 804–810 (2007).

112. A. J. Obregon-Tito, R. Y. Tito, J. Metcalf, K. Sankaranarayanan, J. C. Clemente, L. K. Ursell, Z. Z. Xu, W. Van Treuren, R. Knight, P. M. Gaffney, P. Spicer, P. Lawson, L. Marin-Reyes, O. Trujillo-Villarroel, M. Foster, E. Guija-Poma, L. Troncoso-Corzo, C. Warinner, A. T. Ozga, C. M. Lewis, Subsistence strategies in traditional societies distinguish gut microbiomes. *Nat. Commun.* **6**, 1–9 (2015).

113. A. Shaiber, A. Murat Eren, Anvi'o snakemake workflows (2018), (available at http://merenlab.org/2018/07/09/anvio-snakemake-workflows/).

114. A. Murat Eren, J. H. Vineis, H. G. Morrison, M. L. Sogin, A Filtering Method to Generate High Quality Short Reads Using Illumina Paired-End Technology. *PLoS One*. **8**, e66643 (2013).

115. Y. Peng, H. C. M. Leung, S. M. Yiu, F. Y. L. Chin, IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*. **28**, 1420–1428 (2012).

116. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat. Methods*. **9**, 357 (2012).

117. T. S. Jørgensen, Z. Xu, M. A. Hansen, S. J. Sørensen, L. H. Hansen, Hundreds of circular novel plasmids and DNA elements identified in a rat cecum metamobilome. *PLoS One*. **9**, e87924 (2014).

118. D. E. Wood, J. Lu, B. Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).

119. J. Lu, F. P. Breitwieser, P. Thielen, S. L. Salzberg, Bracken: estimating species abundance in metagenomics data. *PeerJ Comput. Sci.* **3**, e104 (2017).

120. S. C. Watts, S. C. Ritchie, M. Inouye, K. E. Holt, FastSpar: rapid and scalable correlation estimation for compositional data. *Bioinformatics*. **35**, 1064–1066 (2018).

121. G. Csardi, T. Nepusz, Others, The igraph software package for complex network research. *InterJournal, complex systems*. **1695**, 1–9 (2006).

122. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

123. J. Heer, S. K. Card, J. A. Landay, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, NY, USA, 2005), *CHI '05*, pp. 421–430.

124. M. D. M. Leiserson, H.-T. Wu, F. Vandin, B. J. Raphael, CoMEt: a statistical approach to identify combinations of mutually exclusive alterations in cancer. *Genome Biol.* **16**, 160 (2015).

125. M. K. Gibson, K. J. Forsberg, G. Dantas, Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.* **9**, 207–216 (2015).

126. B. P. Alcock, A. R. Raphenya, T. T. Y. Lau, K. K. Tsang, M. Bouchard, A. Edalatmand, W. Huynh, A.-L. V. Nguyen, A. A. Cheng, S. Liu, S. Y. Min, A. Miroshnichenko, H.-K. Tran, R. E. Werfalli, J. A. Nasir, M. Oloni, D. J. Speicher, A. Florescu, B. Singh, M. Faltyn, A. Hernandez-Koutoucheva, A. N. Sharma, E. Bordeleau, A. C. Pawlowski, H. L. Zubyk, D. Dooley, E. Griffiths, F. Maguire, G. L. Winsor, R. G. Beiko, F. S. L. Brinkman, W. W. L. Hsiao, G. V. Domselaar, A. G. McArthur, CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517–D525 (2020).

127. F. Trigodet, K. Lolans, E. Fogarty, A. Shaiber, H. G. Morrison, L. Barreiro, B. Jabri, A. Murat Eren, High molecular weight DNA extraction strategies for long-read sequencing of complex metagenomes. *bioRxiv* (2021), p. 2021.03.03.433801.

128. A. Bankevich, S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham, A. D. Prjibelski, A. V. Pyshkin, A. V. Sirotkin, N. Vyahhi, G. Tesler, M. A. Alekseyev, P. A. Pevzner, SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19**, 455 (2012).

Supplementary Figures

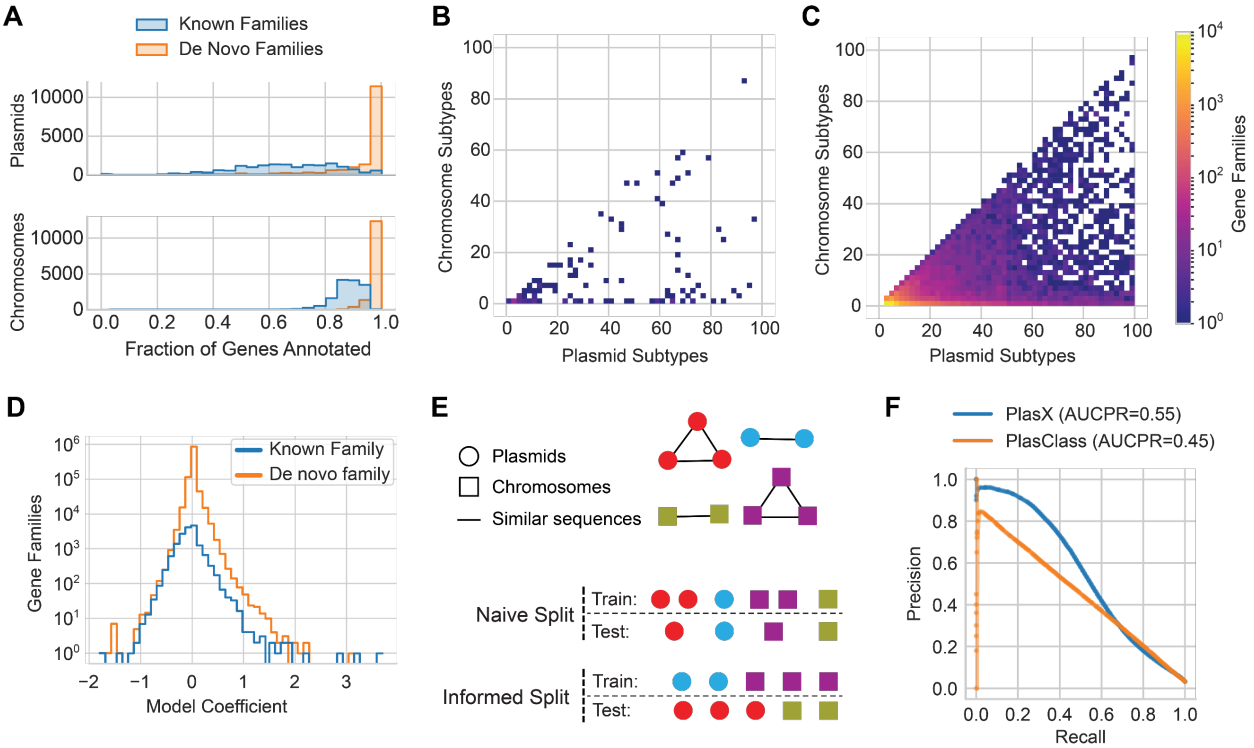

Fig. S1. **(A)** Histograms of reference sequences, based on the fraction of genes that have known or *de novo* family annotations. **(B-C)** Two-dimensional histograms of known (B) and *de novo* (C) gene families, based on the number of plasmid and chromosomal subtypes each family contains. The number of gene families is log-scaled. Only gene families that are enriched in plasmid subtypes (bottom-right triangle) are shown. **(D)** Histograms of the coefficients learned by PlasX. **(E)** Diagrams of different training-test split configurations for cross-validation. A random 'naive' split of plasmids and chromosomal sequences results in training and test sets that have similar sequences, due to the existence of plasmid and chromosomal subtypes that each contain highly similar sequences. An 'informed' split assigns all sequences of the same subtype to either training or test, creating a more representative evaluation of a model's ability to generalize to unseen sequences. Colors and edges represent similar sequences. **(F)** Precision-recall curves using 4-fold cross-validation and a naive split. AUCPR was calculated using a typical, unweighted formula.

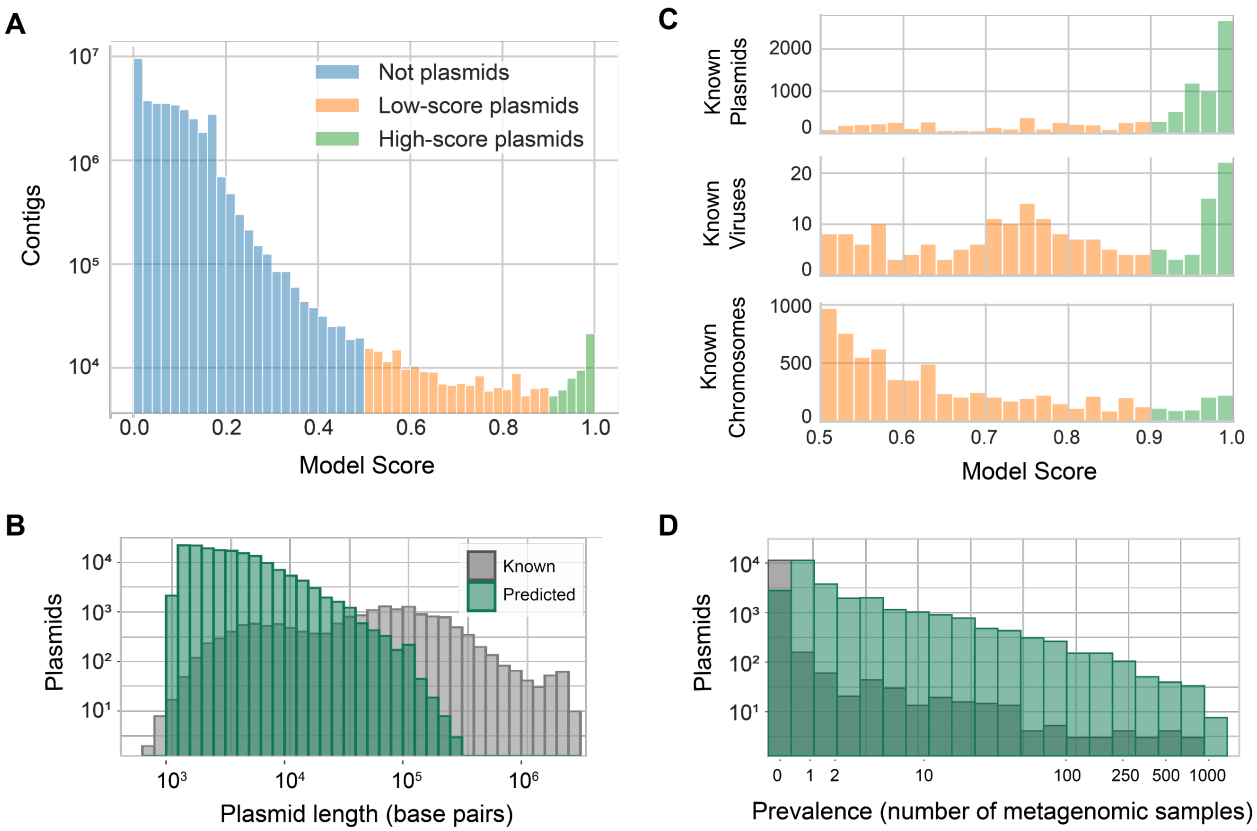

Fig. S2. (A) Model scores of all contigs assembled from all 1,782 metagenomes. 226,194 plasmids were predicted by applying a score threshold of >0.5. Of these, 50,163 plasmids were high-scoring (≥0.9 score). **(B)** The sequence length of known and predicted plasmids. **(C)** Model scores of predictions that matched a sequence in NCBI (≥90% alignment identity and ≥90% coverage of prediction). Predictions are labeled as a known 'plasmid', 'virus', or 'chromosome' based on the presence of these words in the description of the matching NCBI sequence. We searched NCBI for only the filtered set of 100,719 non-fragment predictions. **(D)** The prevalence of known and predicted plasmids across all metagenomes.

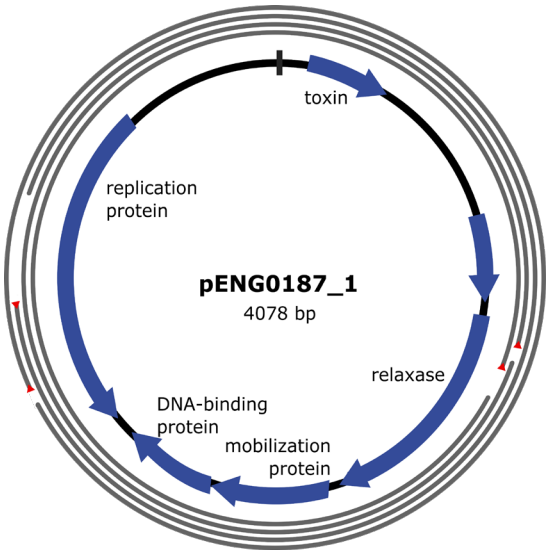

Fig. S3. *B. fragilis* 216 long reads aligned to pENG0187_1, demonstrating circularity. 4 (of 500) reads are shown. Red triangles designate the beginning of a long read.

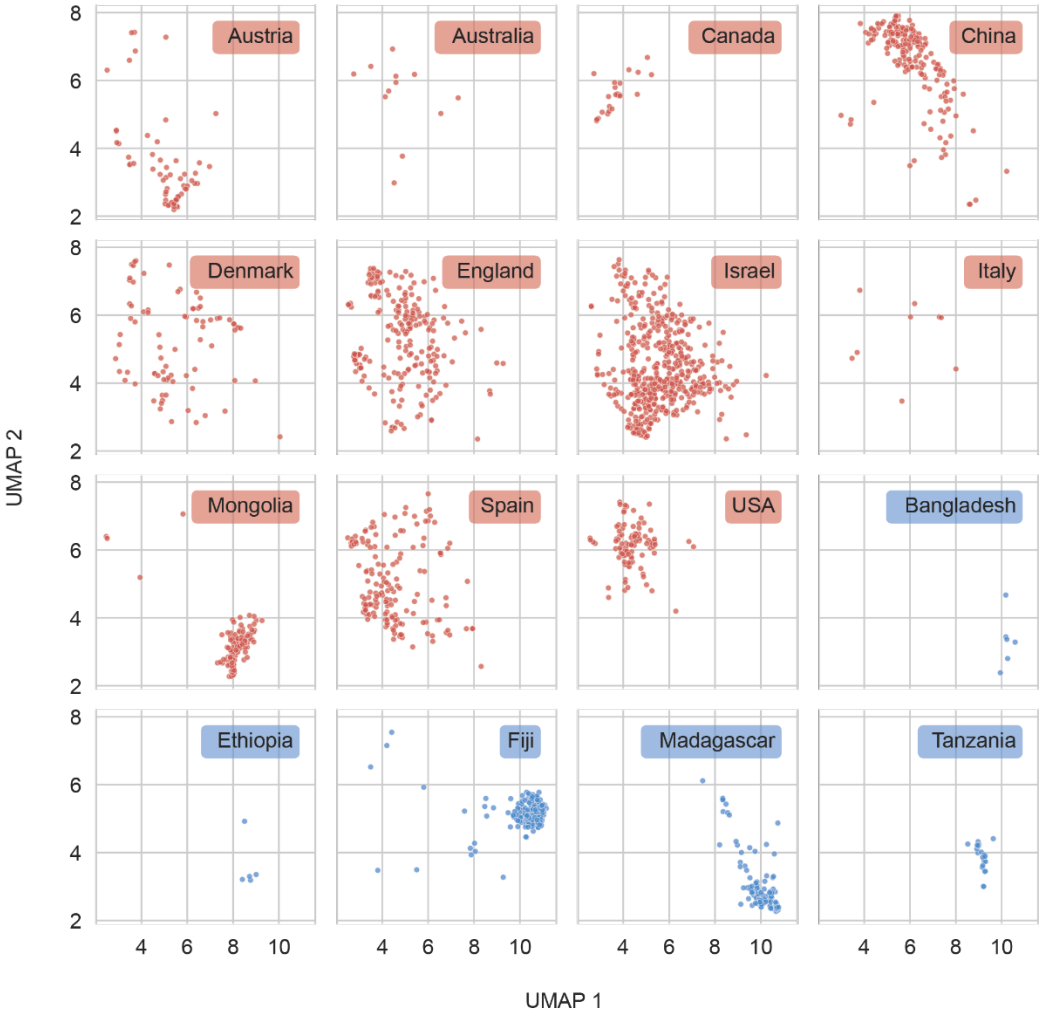

Fig. S4. UMAP plot as in Fig. 4C. Metagenomes have been partitioned to show clustering within each country.

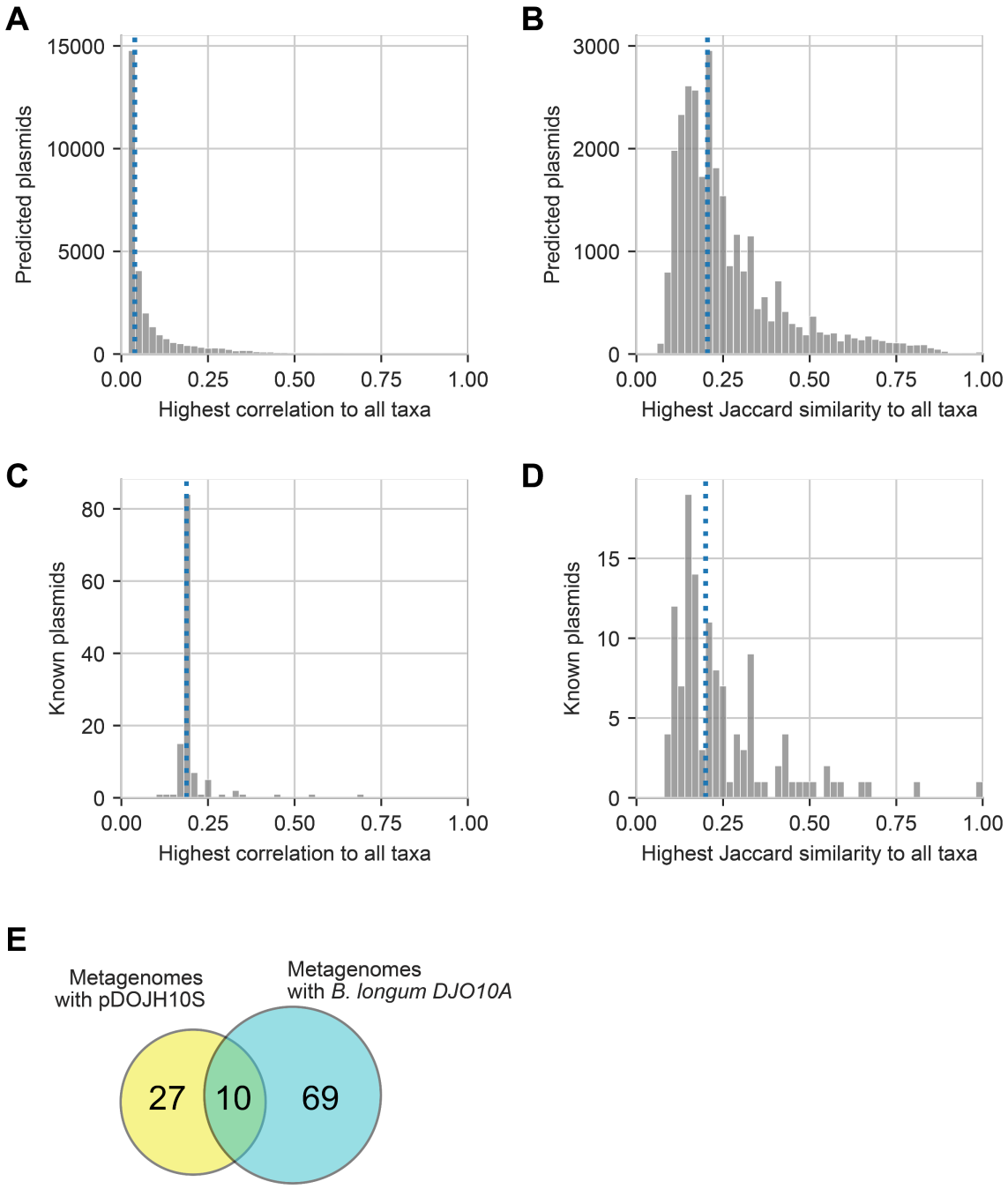

Fig. S5. Comparison of the ecological distributions of plasmids and microbial taxonomy. We measured the association between every plasmid and taxon by calculating the correlation between their abundance levels across metagenomes, using the SparCC technique (*51*). As another association measure, we applied thresholds to the abundance levels and then calculated the Jaccard similarity between the metagenomes containing the plasmid versus those containing the taxon. We restricted analyses to plasmids that were present in at least 5 metagenomes. **(A-B)** For every predicted plasmid, we identified the taxon with the highest correlation **(A)** or Jaccard similarity **(B)**. **(C-D)** We did the same to identify the best matching taxa of known plasmids. Blue lines indicate the median of each distribution. **(E)** Venn diagram

depicting the discordance between the metagenomes containing a plasmid pDOJH10S and those containing its cognate host, a *B. longum* strain.

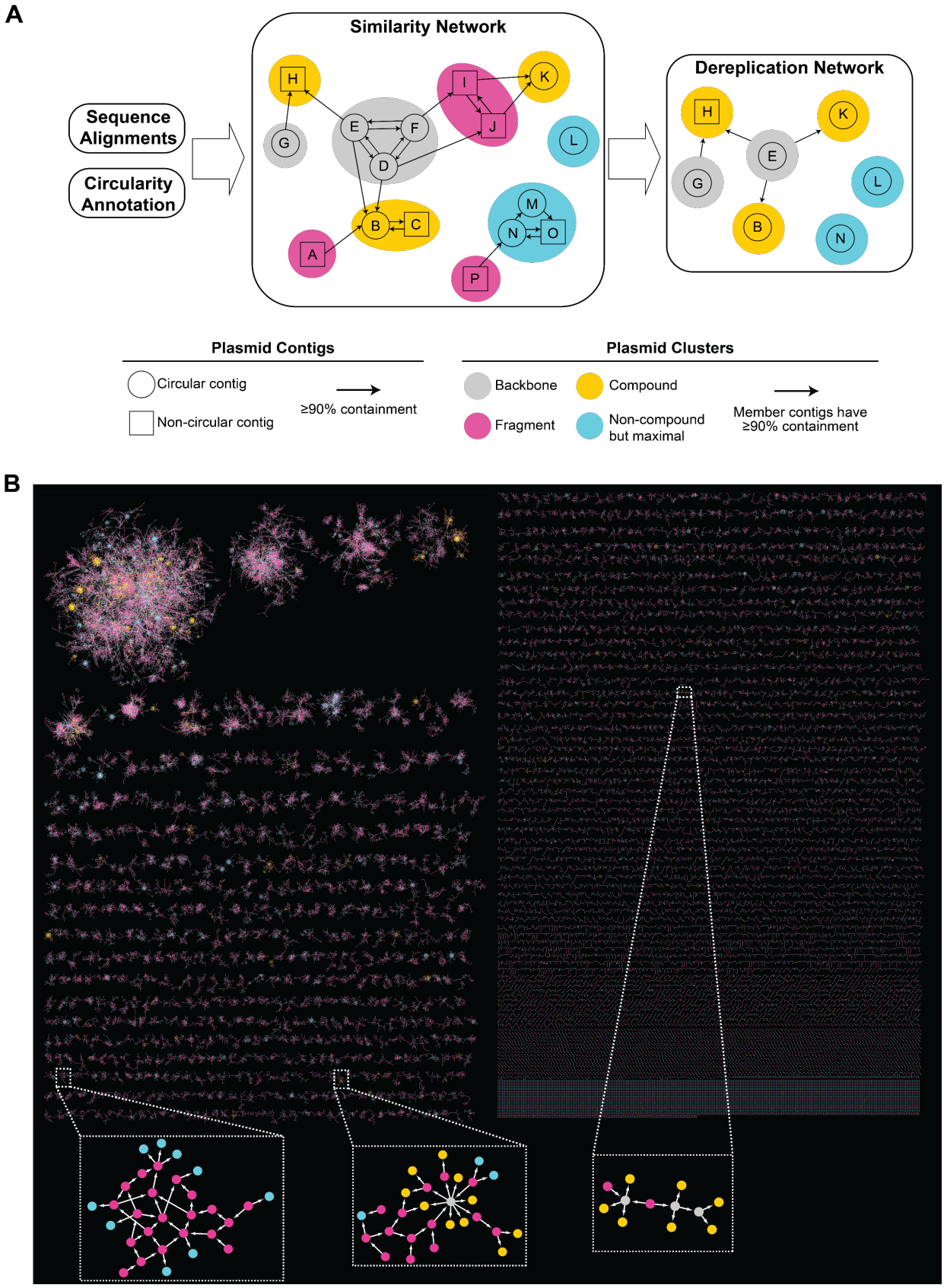
A
Similarity Network
Sequence Alignments
Circularity Annotation
Dereplication Network
H
G
E
F
D
I
J
K
L
B
C
A
M
N
O
P
Plasmid Contigs
Circular contig
Non-circular contig
≥90% containment
Plasmid Clusters
Backbone
Compound
Fragment
Non-compound but maximal
Member contigs have ≥90% containment
B

Fig. S6. (A) Diagram of the MobMess algorithm for dereplicating plasmids and discovering plasmid systems. All-vs-all sequence alignments and circularity information are used to construct a similarity network of plasmid contigs. Similar contigs are clustered, and every cluster is labeled as either a backbone, fragment, compound, or non-compound maximal. A plasmid system consists of a backbone cluster and the compound clusters connected to the backbone. This example shows two systems: one system has G as the backbone (H is the compound plasmid), and another system has D, E, and F as the backbone (B, C, H, and K are the compound plasmids). To dereplicate, fragment clusters are discarded and a representative sequence is chosen for every non-fragment cluster. **(B)** Network of clusters of predicted plasmids. All clusters are shown except those that are not connected to any other cluster.

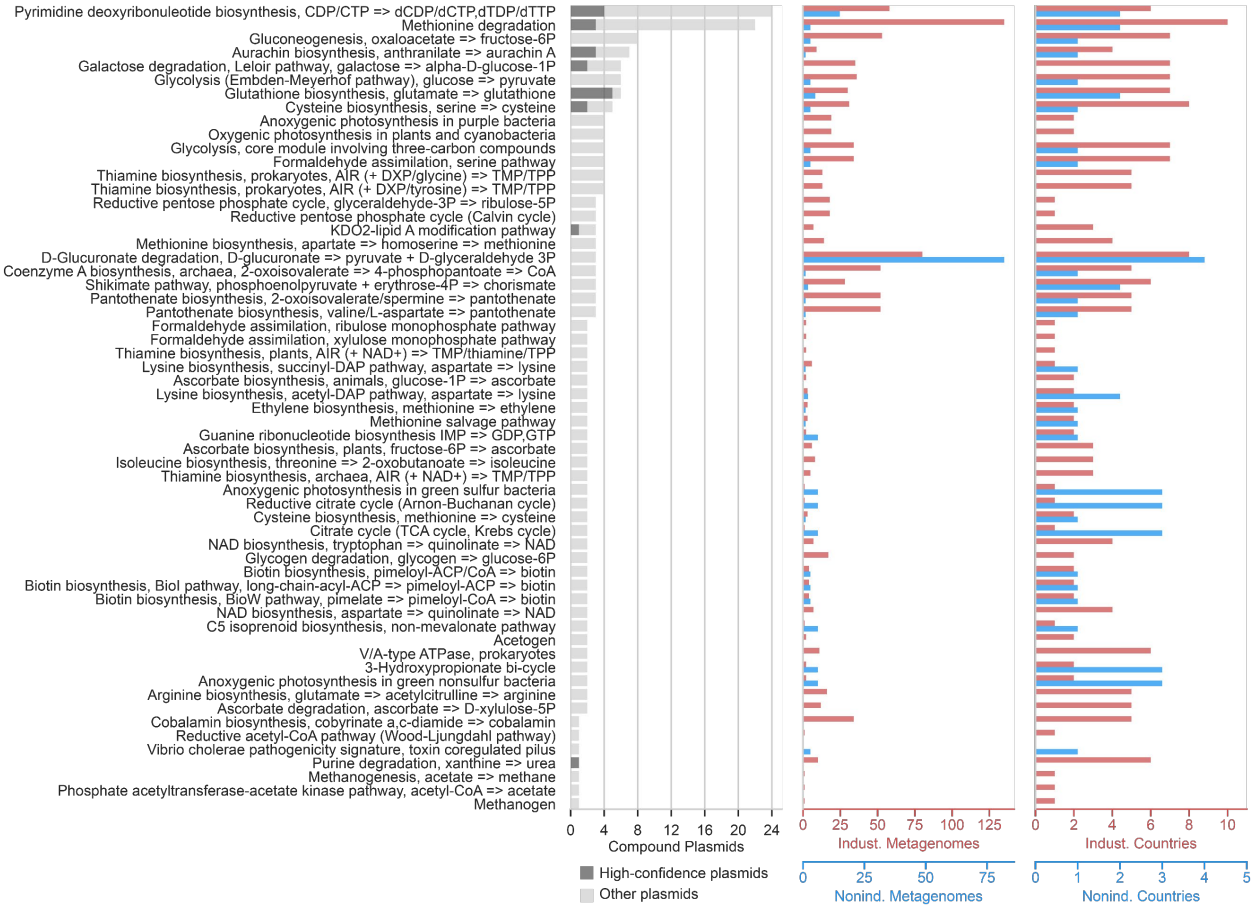

Fig. S7. Functional annotation of cargo genes to KEGG modules, similar to Fig. 6A-B. This plot excludes KEGG modules that occur in only one plasmid system. To avoid redundancy with Fig. 6A, this plot also excludes modules that occur in cargo genes annotated to antibiotic resistance.

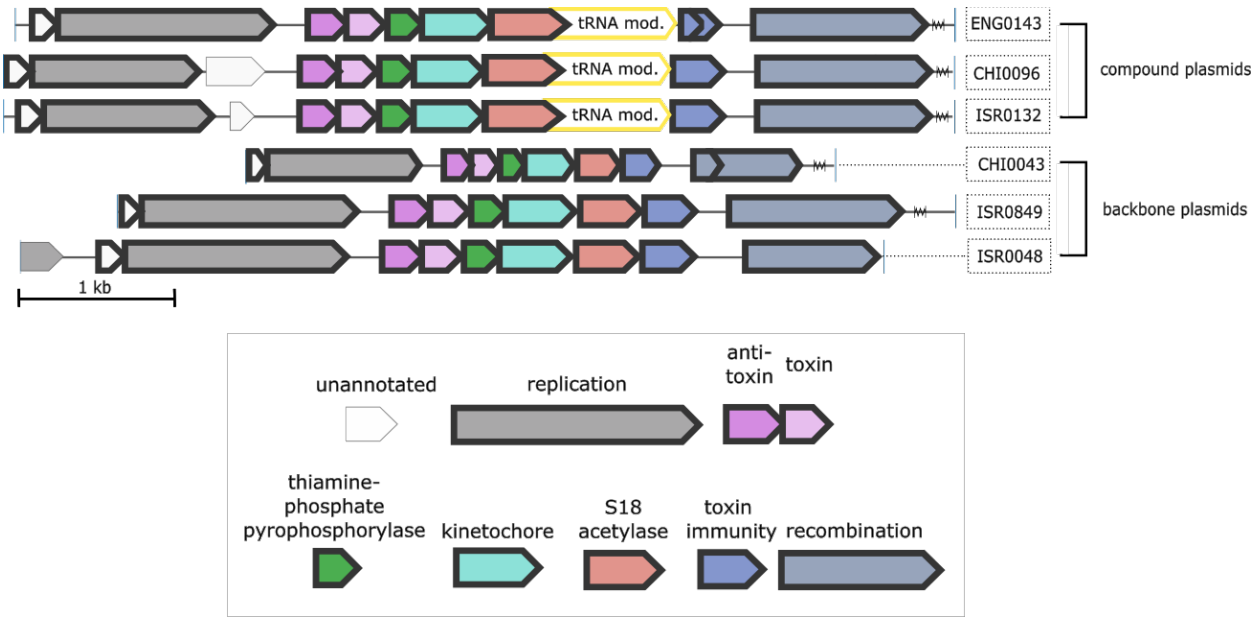

Fig. S8. Plasmid system PS1110. Compound plasmids in this system contain a gene that encodes two enzymes, a tRNA Gm18 2'-O-methylase (yellow, 'tRNA mod.') and a Ribosomal protein S18 acetylase (red). Backbone plasmids contain a similar gene that encodes the S18 acetylase but lacks the tRNA methylase. Backbone genes have a thick, black outline.

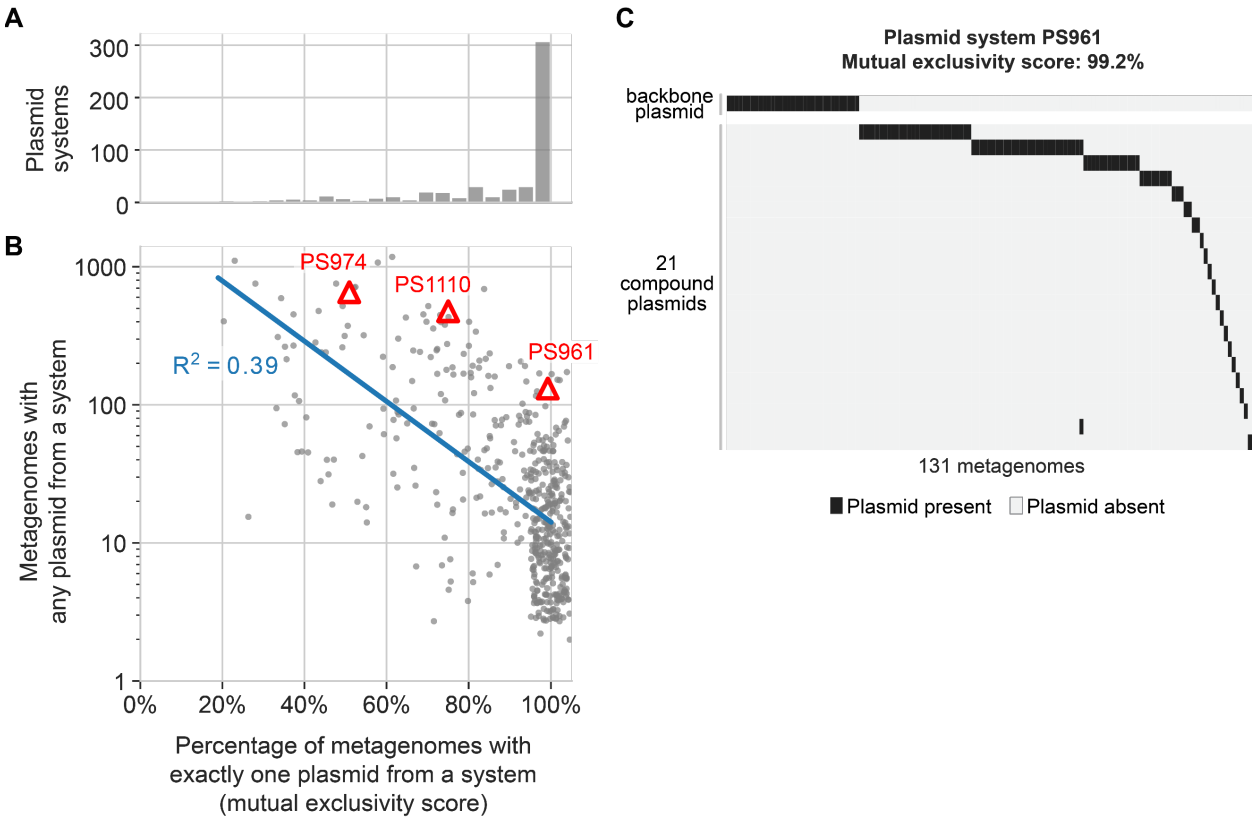

Fig. S9. For every plasmid system with two or more compound plasmids, we defined a 'mutual exclusivity score' to quantify how often its plasmids segregated to different metagenomes. This score is defined as the ratio of metagenomes with exactly one of the system's plasmids to the metagenomes with any of the system's plasmids. **(A)** Histogram of mutual exclusivity. **(B)** The inverse relation between the mutual exclusivity and prevalence of a system. Red triangles represent examples of plasmid systems that are highly prevalent (present in >100 metagenomes) but are more mutually exclusive than expected by a linear regression (blue line). For easier visualization, the x- and y-coordinates of systems with >90% mutual exclusivity were randomly jittered within ±5% of the axes lengths. **(C)** Visualization of mutual exclusivity in PS961, one of the examples in B.

Supplementary Tables

Table S1.

Summary of reference plasmids and chromosomes

Table S2.

Names, accession numbers, and metadata of metagenomes

Table S3.

Summary of predicted plasmids (model scores, orthogonal support, circularity, and NCBI blast results)

Table S4.

Summary of plasmid systems

Table S5.

Gene sequences for plasmid systems shown in Fig. 5D, fig. S8 and Fig. 6F

Table S6.

DNA extraction and sequencing parameters for long read sequencing of isolate genomes